How to Make Achievement Tests

ROBERT M. W. TRAVERS

How to Make Achievement Tests

How to Make Achievement Tests

By ROBERT M. W. TRAVERS
Associate Professor of Education
Teacher Education Division
Board of Higher Education
New York City

THE ODYSSEY PRESS ✣ ✣ ✣ NEW YORK

PRINTED IN THE UNITED STATES

THIRD PRINTING

Preface

THE MAIN PURPOSE of this book is to help teachers develop the types of evaluation instruments that are known as objective tests of achievement. A secondary purpose is to provide teachers with a technique for defining educational goals. Throughout the book emphasis is placed on the fact that objective tests usually measure only a few of the many outcomes of most educational programs. It is hoped that the reader will acquire an understanding of the limitations as well as of the value of objective tests of achievement.

In order to avoid confusion the author has used the term *educational goal* in preference to the term *educational objective* throughout the book. Although this terminology is not generally used, it is adopted in this book because to discuss *objective* tests and educational *objectives* in the same paragraph seemed likely to bewilder the reader. The term *objective test* is used according to current practice in the field of educational measurement.

It must be recognized by the reader that many of the statements in this book are not based on systematic research but represent the opinions of those who have worked on various aspects of the problem of measuring achievement. These opinions need to be tested as hypotheses. Similarly, educators are still in the exploratory stage of finding techniques for defining educational goals in a form which makes the extent to which they are achieved a measurable

function. The techniques presented in this book for defining educational goals admittedly leave much to be desired, but they are widely used and represent a considerable improvement over earlier techniques in terms of their usefulness.

The author gratefully acknowledges the permission given him to reproduce some examples of achievement tests from *The Forty-Fifth Yearbook* of the National Society for the Study of Education. He is also indebted to the Educational Testing Service for permission to reproduce certain questions from their published tests. Sources of good illustrative materials are few, and without the help of these two sources this book would have been much harder to prepare.

The author is indebted to Dr. Helen M. Walker, who read the first draft of the manuscript and suggested many additions which were incorporated in the final draft. Many helpful suggestions were also made by Dr. Edward J. Furst, Dr. Wimburn L. Wallace, and Mrs. Eleanor Ryckman.

The large chart inserted in the back of the volume was prepared by Mr. Joseph W. Leonard and Mrs. Mildred Leonard as part of a test-construction project, and their willingness to have it reproduced here is greatly appreciated.

ROBERT M. W. TRAVERS

New York City
December, 1949

Contents

How to Make Achievement Tests

Chapter One

Introduction

THIS BOOK IS DESIGNED as a guide for teachers in measuring certain achievements of their pupils by means of objective tests. It is not a comprehensive manual on the measurement of the outcomes of education, for many of the most important outcomes cannot be measured with the typical objective examination. Objective examinations are so widely used that teachers should know what outcomes they can measure, what outcomes they cannot measure, and the best method of building such examinations.

MEASUREMENT AND EVALUATION

It is usual to refer to the appraisal of the outcomes of education as *evaluation* because it is part of the process of determining the values that are inherent in the educational process. Teachers are concerned not only with what a person does but also with the worth of his activities. Consequently, teachers are continually making value judgments about the behavior of their pupils and those judgments are referred to

as evaluations. When the teacher assesses student development he determines not only what development has taken place but also whether the development was good and desirable. In order to indicate that the teacher is trying to do more than merely measure development, the process of appraising student progress is referred to as *evaluation* rather than as measurement.

THE EXPANDING MEANING OF ACHIEVEMENT

It is common among psychologists to use the term *achievement* to refer to outcomes related to knowledge of traditional subject matter because the kinds of achievement tests which psychologists have built have been subject-matter tests. However, the term *achievement* has a much broader meaning to the teacher, and refers to the learning of interests, attitudes, appreciations, and social adjustments as well as to the learning of facts. Many elementary-school teachers would consider the development of good social adjustment and good social habits at least as important achievements as the development of good reading habits, and much of the planned activity of the elementary school is directed towards the former goal. Consequently attempts to measure school achievement solely through the use of a few typical objective tests represent an outmoded view of evalution. Currently available objective tests and teacher-made tests are valuable for measuring certain limited aspects of achievement, but they should not be considered to measure a major fraction of the important outcomes of education.

SYSTEMATIC VERSUS UNSYSTEMATIC EVALUATION PROCEDURES

Evaluation procedures are often thought of as trimmings on the educational process or as something which occurs in the last few hours of a course. Nothing could be further from the truth, for evaluation is an inescapable and continuous process.

There are certain inadequacies in the day-to-day evaluations which make it necessary to add systematic evaluation procedures to the unsystematic ones. The systematic evaluation procedures may require the use of tests, check lists, and other instruments, and they supplement rather than replace the informal procedures. Informal evaluation procedures often lack both reliability and validity, and thus often give either false impressions, or impressions which are largely influenced by the hopes and wishes of the teacher. Also, they tend to be based on selected observations. The teacher is more likely to remember actions of pupils indicating the achievement of desirable goals than those indicating that the same goals have not been adequately achieved. In this connection Darwin once remarked that he tended to forget facts which did not agree with his theory and remembered those that provided supporting evidence. Therefore, it is both desirable and necessary for teachers to devise systematic methods of appraising student development.

Another reason for using systematic evaluation procedures is that the pupil may not show in his behavior either evidence indicating the achievement of a desired goal or evidence indicating that the goal has not been achieved. As a matter of fact, in most situations the teacher will not have evidence of the extent to which most goals are achieved in

most individuals unless systematic evaluation procedures are used. Even a teacher who has been in daily contact with twenty-five children for a period of a year may still have a very inadequate basis for appraising the development of each child unless systematic evaluation procedures are used. Until classes are reduced in size to a point far below present-day levels and probably below levels set by practical politics, systematic evaluation procedures will be essential for appraising pupil development.

TEACHER EVALUATIONS AND PUPIL EVALUATIONS

Certain educators who belong to the so-called "left wing" of the progressive education movement have stated that, in their ideal school, the most important part of the evaluation procedure would be the evaluation which a student makes of his own performance. This point of view assumes that the student knows, first, what objectives are to be achieved, and second, whether he has achieved them. In certain fields it is possible for the student to know just what is to be achieved and consequently he can have a clear idea of the anticipated outcomes and be able to match and compare actual outcomes with expected outcomes. For example, a student who is attempting to make a small vase of clay can compare his product with an actual model, or a student who is trying to make certain calculations may be able to test by empirical means the accuracy of his computations. On the other hand, in many fields it is quite impossible for the pupil to determine on the basis of his own judgment whether he has achieved the goal he was supposed to achieve. In the area of attitudes, it is quite evident that the student is often one of the poorest judges of actual attitude change. The teacher

may struggle to make the attitude of the adolescent more liberal in matters related to behavior towards minorities, but most students will claim a liberal attitude both before and after the experiences provided by the teaching situation. Similarly, if children are learning to write compositions which are designed to transmit ideas to other people without distortion, it is evident that the best evaluation of such a composition is in terms of the reaction of the individual to whom it is supposed to transmit certain given ideas. A child's composition, written with a purpose in view, should be judged to a considerable extent in terms of the degree to which it achieves that purpose, and not in terms of the satisfaction which the author feels about his composition.

While it is agreed that the student should learn to make evaluations of his behavior, this does not mean that the student's own evaluations should be the only ones made except in very unusual circumstances. One of these circumstances occurs when the primary outcome of a course is the enjoyment experienced by the student. Many courses in art and literature are of this character and the success of such courses should be measured primarily in terms of the satisfaction which the students derive from them. Under such circumstances, it is very evident that the evaluation by the student of his own enjoyment forms the only satisfactory evidence of the extent to which the goals of the course have been achieved. However, such situations are rare.

In a democratic society, behavior is evaluated largely in terms of judgments by the members of that society. Thus, the evaluations of the individual child made by the teacher, the parents, and the classmates are all important in appraising the behavior of the child.

TECHNIQUES OF SYSTEMATIC APPRAISAL

It is quite unfortunate that it is still widely believed that systematic appraisals of achievement can be discussed comprehensively under the heading "examinations." The reason is that for generations the only available instrument for the systematic measurement of achievement was the essay examination, and it was the teacher's stock-in-trade device for all marking and grading. The traditional type of essay examination had some merit as an evaluation device so long as teachers were concerned with developing knowledge and understanding in limited subject-matter fields, but it has become progressively more inadequate as an evaluation device as the goals of the curriculum have expanded. It is fairly obvious that, in many cases, questionnaires can provide much better evidence of the student's development than the usual type of essay examination. It is also conceivable that outcomes related to health education can be measured more adequately by finding out the number of children who come to school with colds or by observing the food selected by the children in the cafeteria than by giving them an examination on their knowledge of good health practices. Evidence of the achievement of the outcomes of education should be sought wherever it can be found and teachers should not be bound to any particular type of appraisal procedure.

INADEQUACIES IN THE STATEMENT OF EDUCATIONAL GOALS

Evaluation procedures are meaningful only after educational goals have been clearly stated. The main reason for

the absence of good evaluation procedures in so many schools is that the faculty have not yet stated in precise terms what outcomes they are striving to achieve. Ambitious groups which attempt to develop new curricula and to improve the program provided by a school often fail because they are unable to define in the first place what outcomes are to be achieved.[1]

In fields other than education, attempts to measure the success of an enterprise are part of the normal routine and are taken for granted. The success of a business enterprise is measured largely in terms of a balance sheet, and every student in business school learns the technique involved. The success of a public-health service is measured in terms of the changes in the disease rates and death rate in the community. The success of a fire-prevention campaign is measured in terms of the change in the number of calls for the fire department. But the success of education is usually assumed rather than proved and is estimated in terms of judgment rather than in terms of measured results.

The main reason for this state of affairs is that while the aims of a health department, a traffic-control system, or a fire-department campaign are clear and well defined, and goals of education are usually obscure and couched in terms so vague that they mean different things to different people.

[1] The main concern here is with primary outcomes which have a widespread and, it is hoped, lasting effect on behavior and which operate in regions beyond the domains of behaviors specified by the content of the course. It is true, however, that all the way through a plan of organized study there are subsidiary objectives which must be achieved before the primary objectives can be achieved. For example, in order for the radio technician to be able to repair a receiving set he must first learn about the essential characteristics of electricity. The latter may be referred to as secondary outcomes, the former as the primary outcome. Every experience provided for the student achieves both immediate goals and also some progress towards larger goals. The primary goals are ultimately achieved, if they are achieved, through a great diversity of experiences, while the subsidiary goals are immediate outcomes of individual experiences.

The first step in the appraisal of an educational program is the precise definition of the goals which that program is attempting to attain.

Some educators commonly talk about "successful teaching" as if teaching could rightly be called successful or unsuccessful without definition of the purpose which the teacher is trying to achieve. A naval architect who designs a ship would not call his ship a successful or an unsuccessful ship, though he would be justified in saying that his ship was successful for one particular purpose and maybe unsuccessful for another. A particular ship might have a highly successful design if it were to be used as a river-going freighter, and it might be highly unsuccessful if it were to be used for ocean transportation. Similarly, a teacher may be highly successful so far as his aim is to provide a series of pleasant hours for the students, but highly unsuccessful so far as his aim is to develop a certain degree of understanding of and insight into the workings of our society. It is meaningless to talk about the over-all success of teaching, since success can be measured only so far as the teaching is directed towards a definite purpose. A naval architect would never be satisfied to say that he was trying to build a ship which would be seaworthy, elegant in appearance, small enough to enter most ports, sufficiently fast, and capable of being managed by a relatively small crew. Such specifications would be quite useless. Yet, in education one usually finds sets of goals which are equally vague. The teacher may aim at developing students who are critical in their ways of thinking, able to interpret data, socially sensitive, able to appreciate the deeper values of life, well adjusted socially, and with other desirable qualities. These goals are worthy enough, but they are couched in such vague terms that for practical purposes nobody can ever tell whether they have been achieved. The teacher who aims to achieve social

sensitivity in his students, like the naval architect who wants simply to build a good ship, does not know what he is doing.

But the naval architect finds it a simple matter to define his goals with precision. He can specify the size of the ship to be built, in terms of tonnage displacement; he can specify the speed at which the ship is to run; he can calculate the engine size necessary for the ship to do what it is supposed to do; he can allow a given number of square feet of deck space; he can specify the exact number of cubic feet which the hold of the ship is to contain; and he can define precisely the dimensions and other characteristics of the ship. He can then easily check whether the final product meets these specifications. The teacher, however, does not yet have a recognized way of specifying the outcomes of the educational process as precise as the methods used by the naval architect, the businessman, or the engineer. It should be noted that the usual dictionary type of definition is not adequate for defining educational goals. Indeed, for this purpose the traditional dictionary definition has been largely discarded by educators.

METHODS OF DEFINING EDUCATIONAL GOALS

Examine a goal such as "the ability to interpret data," towards which teachers in many fields are striving. This goal is thoroughly vague, too vague to be of practical use, but the idea which it conveys can be clarified by listing examples of the things which a student may be expected to do if he has achieved it. Some of the things which one teacher hoped students would be able to do included the following:

1. Describe in general terms the trend shown by a graph.
2. Read the values of a given point on a graph.

3. Interpret a graph as a rate of change.
4. Explain the meaning of the area under a graph.
5. Describe in general terms the data presented in a table.
6. Identify relationships between variables.
7. Estimate missing values in a table.
8. Identify assumptions made in drawing certain conclusions from data.

It should be noted that the teacher who defined the ability to interpret data in terms of these items had in mind mainly statistical data. Anyone who had heard this teacher say that he was trying to develop the ability to interpret data might have attached an entirely different meaning to that statement. It might have been understood to mean the ability to interpret detailed descriptions of the types which are found in geology and biology textbooks.

The items of behavior in the list above are commonly referred to as evaluative criteria, which means that they can be used to determine whether the student has achieved what he was supposed to achieve. Once the list of evaluative criteria has been developed, it is possible to appraise the student's achievement by placing him in situations in which he may respond with one or more of these specific behaviors. For example, he may be given a table of figures and asked to estimate missing values, or he may be given a graph and asked to read the values of a given point on it.

The illustration above shows that one way of defining an educational goal so that all may know just what it means is to list some of the things which a student may do if the goal has been achieved in him.[2]

[2] The basic difficulty in defining educational goals is due to the fact that psychologists have not yet developed a classification of human behavior which is useful for this purpose. A comprehensive taxonomy of human behavior which had some numerical value assigned to each category of behavior would simplify the educator's task. It would also provide teachers with a common language for discussing educational goals and ensure that those who used the same terms referred to the same concepts.

The evaluative criteria form a basis for appraising student progress. They also serve the purpose of indicating precisely what is meant by the goal which has been stated in general terms. The same process of defining educational goals may be used when they are not concerned with knowledge but with interests, attitudes, or appreciations. For example, one twelfth-grade teacher wished to develop in his students an interest in current affairs; the following are the evaluative criteria which he listed in planning his course:

1. Listens to news reports on the radio.
2. Reads some of the news items as well as the "funnies" in the newspaper.
3. Spends some time each week with a news magazine.
4. Recognizes how certain news items affect his future.
5. Looks for the results of elections.
6. Enters into discussions of current events.
7. Knows the names of some of the men in local, state, and federal government.
8. Offers opinions on current affairs.

If the student did some of these things, it would be reasonable to suppose that he was showing some interest in current affairs and that the goal had been at least partially achieved.[3]

Another teacher was interested in developing the capacity for "healthful and successful family living." This goal can be interpreted in a great variety of ways, but becomes fairly

[3] One basic pedagogical difficulty arises out of the definition of educational goals in terms of specific behaviors. It sometimes happens that teachers misunderstand the purpose of listing specific items of behavior, and as a result attempt to provide training in the specific behavior elements rather than training in the general domain of behavior which the specifics represent. Under such conditions, the students may manifest the desired specific behaviors at the end of a course, but may fail to manifest other specific behaviors which are representative of the same domain. In this latter case, appraisal of student development can be undertaken only by testing for the occurrence of behaviors from the domain defined by the goals but other than those actually listed in defining the goals.

clear when it is considered in the light of the following fairly specific behaviors which the teacher used for defining it:

1. Keeps own room clean.
2. Cleans the bathtub after use.
3. Isolates self when ill.
4. Makes no loud noises when others are resting.
5. Looks to parents for guidance.
6. Treats parents with courtesy.
7. Does his share of the chores around the house.
8. Does not keep family waiting for meals.
9. Volunteers to help family with special chores that arise.
10. Does not try to get special privileges which siblings do not get.
11. Shares gifts of candy, etc., with the family.
12. Volunteers to do the chores of other family members when they are sick.
13. Suggests ways of saving money for the family.

Here again the evaluative criteria help to provide a clear picture of the kinds of things that the teacher is trying to develop. In this case it should be noted that it would not be possible to appraise the student's achievement of this goal by means of any of the common types of paper-and-pencil examinations. The only reasonable method of determining whether this goal had been achieved would be to find out whether the boys or girls manifested any, some, or all of these behaviors in the home situation. Paper-and-pencil tests in such a case would indicate only whether the student knew the kinds of things he should do in his home and not whether he would actually do them. In this case as in many other cases there is likely to be a discrepancy between the verbal response to a situation described in print and the kind of response made when the person is faced with the same situation in daily life.

The examples above have been discussed in order to illustrate how educational goals become meaningful concepts by being defined in terms of the behaviors which will be accepted as evidence of the achievement of the goals. Teaching should not be started nor should the systematic evaluation of student progress begin until the goals of teaching have been properly defined in this way.

In summarizing this section, it may be said that the operational method of defining goals discussed above has two advantages. First, it supplies a clear concept of what is meant by each goal. Second, it supplies a basis for determining the extent to which the goals are achieved.

Chapter Two

Steps in Planning Evaluation Instruments

THE BASIC PLAN of an achievement test is referred to as the blueprint. The term is derived from engineering practice, in which an object to be manufactured is first pictured by a series of drawings which are often reproduced by the blue-print process. These drawings specify exactly what is to be made and any properly equipped shop can make the object thus specified. In engineering it makes little difference who makes the object specified in the drawings, provided the object is made according to specifications. While the engineer's drawings are exact specifications of what is required, the blueprint for an achievement test represents only an attempt to draw up precise specifications. If two teachers started building a test from the same blueprint, they would develop slightly different instruments because they would interpret some aspects of the test plan in slightly different ways. However, it is much better to start with the best test plan that can be made than to start with a poorly organized test plan. The following pages describe the steps necessary for executing one method of planning achievement tests.

STEP I. STATING THE EDUCATIONAL GOALS IN GENERAL TERMS

The first step in planning an achievement test is to state all the educational goals in general terms.[1] For practical purposes it has been found that from eight to fifteen statements of separate goals will form a satisfactory starting point. Fewer statements will mean that, if they are to cover the goals of the program, they will have to be too broad in scope for convenience. A large number of broad goals usually constitute a rather cumbersome structure with which to start any educational program.

Teachers are familiar with lists of goals stated in general terms. The technical literature of education is filled with such lists drawn up by various bodies of "experts." Three examples of teacher-made lists are given below. They are not presented as models, but rather as the typical raw materials with which the teacher starts planning a course.[2]

EXAMPLE I. GOALS FOR A COURSE IN CONSUMER MATHEMATICS

1. The ability to handle the quantitative aspects of daily life

[1] While it is recognized that the problem of selecting educational goals is an important and complex one, it cannot be considered here. A summary of the major methods of arriving at goals would fill a volume by itself. If the reader wishes to investigate the problem, it is suggested that he turn to one of the well-known works on curriculum design. Almost any such work will discuss the various methods of selecting educational goals and the assumptions underlying each of these methods.

[2] Readers should not consider these goals either as validated or as necessarily subscribed to by the present writer. They were developed by teachers working on evaluation problems and were based to an unknown extent on published lists of goals which have been validated in some way. They should not be used as teaching goals merely because they happen to be printed in this book.

2. The ability to spend money wisely
3. A knowledge of sources of reliable information concerning goods and services
4. Skill in evaluating goods and services
5. Ability to weigh values in buying on the installment plan
6. Skill in making budgets and in keeping records of receipts and expenditures
7. Ability to read bills intelligently
8. Skill in reading, interpreting, and constructing graphs
9. Skill in making accurate measurements and in estimating amounts
10. A knowledge of taxation
11. A knowledge of insurance as protection against the hazards of life

EXAMPLE II. GOALS FOR A COURSE IN SHOP DRAWING

1. An appreciation and knowledge of the importance of mechanical drafting in our civilization
2. The ability to express ideas through the use of the language of mechanical drawing
3. A knowledge of the use of drawing in shop and engineering work and the function which such drawing performs
4. A knowledge of the various methods of making prints from tracings
5. Habits of neatness and accuracy in the making of drawings and in the use of instruments
6. An appreciation of well-executed drafting work

7. A desire on the part of the student to demonstrate exceptional ability in this area
8. The ability to interpret simple blueprints of working drawings
9. Skill in the fundamental drafting skills
10. A knowledge of opportunities and working conditions in the drafting trade

EXAMPLE III. GOALS FOR A GROUP-GUIDANCE COURSE

1. Insight on the part of the student into his aptitudes, abilities, and interests
2. A broad knowledge of occupational possibilities
3. An appreciation of the good use of leisure time
4. The habit of searching out pertinent information upon which to base decisions concerning the choice of an occupation
5. An appreciation of the contribution of health and personality traits to occupational success
6. The knowledge of how to obtain a position and how to advance in it
7. Good habits of study
8. A wholesome attitude towards the problems of employers and employees

The educational goals listed above represent rather arbitrary groupings of behaviors which are useful as landmarks in the learning process. The number of goals stated in the general terms used above is quite arbitrary. Any one item in the lists could be broken down into two or three more specific goals. Similiarly, in many cases it would be

possible to combine several of the items into a single and more general goal.

It should also be noted that many of the goals stated in general terms are rather vague. The last statement in the third example is rather typical of the kind of goal which is meaningful only to the teacher who wrote it.

Lists of expected outcomes are too often limited to the acquisition of knowledge, although the teacher is also concerned with the development of attitudes, interests, appreciations, thinking skills, and other outcomes. The teacher should be sure to check his list of goals to see that it includes those of the latter kind, provided they are relevant.

It sometimes happens that when teachers are asked to list goals they list the content of the curriculum. A teacher of history may indicate that a knowledge of American history is the only goal of teaching in that field. Most teachers, however, look upon the content of a course as a means to an end rather than as an end in itself. The purpose of teaching American history goes far beyond the rote learning of facts and dates and it is commonly believed that this course may have important effects upon a pupil's behavior throughout his life. Similarly, it is widely accepted that arithmetic is taught, not for the purpose of enabling the child to manipulate abstract numbers, but to equip him to solve the problems which he encounters or will encounter in the routine of daily life. What is commonly called mastery of the subject matter represents an immediate and usually tempting goal which should be replaced by other goals that have greater significance in the student's present and future life.

The content of a course is usually a means of achieving significant goals, and except for certain special cases it is not an end in itself. It functions like a tool for performing a particular job, but the result of using the tool is different from the tool itself.

STEP II. DEFINING THE EDUCATIONAL GOALS IN SPECIFIC TERMS

The first step—stating goals in general terms—is too often the step where systematic educational planning ends. It has already been pointed out that if general goals are to represent well-defined landmarks towards which the teacher is to direct his efforts, it is necessary that their meaning be clarified. This is done by listing a sample of specific behaviors which would characterize the student in whom those goals had been achieved.

It is very evident that the teacher usually cannot list all of the things which a child may do as a result of achieving a certain goal, since these behaviors are multitudinous; so the problem arises as to how many behaviors should be listed. The solution to this problem is to list as many behaviors as are necessary for giving a clear concept of the way in which pupils will behave if the particular goal is achieved.

In some cases it is both possible and desirable to state all of the specific behaviors that characterize the person in whom an educational goal has been achieved. A course for training janitors might attempt to teach the would-be janitors exactly what to do in a limited number of situations. The aims of the course might include all the specific behaviors which the course is designed to develop. The same would be true of all programs designed to train men and women for simple occupations. However, in those educational programs which attempt to develop complex thinking skills the specific behaviors which may result from the achievement of a particular goal are innumerable. A pupil in a physics course who learns and understands the funda-

mental equations of moving bodies can apply them in an unlimited number of situations. All courses which attempt to develop an *understanding of principles* seek to influence behavior in a great variety of situations. In defining the goals of such courses it is impossible to list all the specific outcomes. Under such conditions it is feasible only to list a sufficient sample to illustrate clearly the goal towards which teaching is directed.

What is the teacher to conclude when he finds that he is unable to list specific behaviors as a means of defining a goal? The answer is simple. When the teacher does not know how the achievement of a particular goal affects the pupil's behavior, the goal is a meaningless thing. Unfortunately, many goals which seem to be worth while on the surface turn out to be quite meaningless when they are subjected to the kind of analysis necessary for adequate definition.

In summarizing the foregoing discussion, it may be said that when this second step in the test plan has been completed, the teacher will have drawn up a set of general statements of educational goals and will have defined each in terms of the specific behaviors which characterize the person in whom it has been achieved.

STEP III. ASSIGNING WEIGHTS TO THE GOALS

It is very evident that all educational goals are not equally important. In any given course some of the goals will be considered by the teacher to be so important that every effort will be made to achieve them in all students, while others will occupy less of the time of the teacher and pupils and possibly be achieved in fewer pupils.

In developing the plan for a course the weight to be attached to each goal should be specified. The simplest method of specifying the weight is to employ numerical terms. One simple system is to distribute 100 points over the goals, giving the greatest number to those that are to receive the greatest emphasis. There is no basis which is entirely satisfactory for assigning these numerical weights to goals, but the simplest system is probably to weight them in terms of the time to be devoted to their achievement. This does not take into account the fact that important goals may be achieved in a short space of time. However, there is a general tendency in education to spend the most time in achieving the goals that are most important and zero time on those that have zero importance.

The assignment of numerical weights to the educational goals is important if an over-all evaluation is to be made of the student's achievement. For example, a student takes a high-school course in chemistry. He would have a good basis for complaining about the evaluation procedures if the teacher appraised his progress entirely in terms of a test of his knowledge of the technical vocabulary of chemistry, for such appraisal would imply that a knowledge of vocabulary was the only important outcome. If the teacher considered vocabulary as only one of the minor outcomes, it should play only a small part in the total appraisal of the individual's achievement. In any test used for evaluation purposes, the outcomes covered should be emphasized in the same degree in which they are emphasized in the course itself.[3] Goals

[3] For those who are mathematically minded, the basic idea in this paragraph may be stated in a symbolic form. If the goals of a course are given weights of W_1, W_2, W_3, . . . W_n, and if the measures of the extent to which an individual achieves these goals are reported in standard scores and are X_1, X_2, X_3, . . . X_n, then an over-all evaluation of the student's achievement would be given by $W_1X_1 + W_2X_2 + W_3X_3 - - - + W_nX_n$.

given equal numerical weights in the course should be given equal weights in the evaluation procedure. In practice, the weight given to each goal measured by an objective test usually, but not necessarily, depends on the number of items devoted to it.

STEP IV. OUTLINING THE CONTENT OF THE COURSE

By the content of the course is meant those experiences which lead to the achievement of educational goals. A study of racial relations in the community may be used to develop desirable attitudes towards other racial groups. In this case, the study of racial relations represents the *content* and the desirable attitudes represent the goals. The *content* is the means of achieving the goals. A series of scientific discoveries may be studied in order to develop in the student the ability to take an analytical and scientific approach to the problems he encounters. The knowledge which the student acquires of the specific discoveries is only a means to an end and not an end in itself, and a study of any ten discoveries might be as useful as the study of any other ten. The goal in this case is to acquire thinking skill by learning how scientists think.

In most academic fields of study there is a fundamental distinction between the content and the goals or outcomes which mastery of the content achieves. However it has already been pointed out that in courses designed to train people in very simple skills the relation between content and goals may be a close one.

There are at least three reasons why an outline of the content of a course is desirable, in most cases, for planning

evaluation procedures.[4] First, if a course has been outlined unit by unit, it is possible to identify the goals which should have been achieved at each stage. Second, in most evaluation procedures it is desirable to study the behavior of the student, not only in problem situations similar to those which he has encountered in the course, but also in problem situations which are relatively novel but which are handled by a procedure similar to the one learned. It is most important to control the amount of novelty in a test situation and it can be controlled only if the content of the course is known. Third, in some courses knowledge may be an important outcome and in such cases the outline of the content of the course summarizes the knowledge that is to be acquired.

The content of a course is usually most easily summarized in terms of units of study. It is common for teachers to organize content in that way. For example, a certain ninth-grade social-studies course was divided by one teacher into the following units:

1. The citizen and his community
2. Welfare agencies
3. Government
4. Social problems in a democracy
5. Economic problems in a democracy
6. Vocational problems
7. Aids for student success
8. Current events

However, a list of units of study is inadequate as a basis either for planning courses or for planning evaluation procedures, because for these purposes it is necessary to break

[4] For those who advocate leaving the planning to the participating students, this step must be carried out at a late stage in the course. However, such a procedure does not exempt the teacher from keeping a careful record of the experiences of the students so that he may later discover whether certain specific experiences achieve certain specific goals.

down the units into smaller component parts—in other words, to describe them in greater detail. The following is an example of a health-education unit on communicable diseases which has been outlined in sufficient detail for test-construction purposes:

UNIT ON COMMUNICABLE DISEASES [5]

A. Causes of infectious disease
   1. Microorganisms
   2. Germs
      a. Nature of germs—kinds, shapes, etc.
      b. Manner of transmission
      c. Diseases transmitted by germs
   3. Filterable virus
      a. Nature of virus
      b. Manner of transmission
      c. Diseases transmitted by virus
   4. Animal parasites
      a. Invisible
      b. Diseases caused by invisible parasites
      c. Visible—lice, fleas, ticks, and mosquitos
B. Control of infectious diseases
   1. Immunization
      a. Acquired
      b. Natural
   2. Vaccines to provide immunity
   3. Vaccines used as tests
   4. Serums to provide immunity
      a. Nature of serums
      b. Serums available
   5. Quarantine
   6. Disinfection

[5] This unit is *not* presented as a model selection of subject matter. It was prepared by a teacher in connection with a particular course and does not necessarily include the material which experts would include in a high-school unit on this topic.

This example illustrates the kind of outline of the content of each unit that it is desirable to make both for planning a course and for planning evaluative procedures. Estimates should be made of the percentage of the total time to be devoted to each unit.

STEP V. THE PREPARATION OF THE BLUEPRINT

When the goals of a course and its content have been defined, the blueprint for the evaluation process may be prepared. In the back of this book an example is given of a blueprint for a test in consumer mathematics. The general goals are entered in the cells along the left-hand edge of the sheet. The various units of the course through which the goals are achieved are entered along the top of the sheet. Examples of specific behaviors which are expected outcomes are entered in the proper cell below each unit and opposite the appropriate goal. The blueprint shows the relationship of content to goals and specifies the domain of behavior within which evaluations are to be made.[6]

STEP VI. THE USE OF THE BLUEPRINT IN PREPARING EVALUATION INSTRUMENTS

Once the blueprint has been drawn up, the next step in the evaluation process is to develop problem situations in which the student will behave in the ways indicated on the

[6] The process of enumerating behaviors within a given domain merely defines that domain of behavior. The teaching process itself is directed towards the development of the particular domain of behavior and in order to be effective must correctly hypothesize mechanisms or processes which underlie and integrate the various behaviors that it is desired to develop.

chart if he has achieved the desired educational goal or will manifest some other form of behavior if he has not achieved the desired goal.

Let the reader assume for the present purposes that systematic evaluation is to be limited to an objective test given at the end of the semester. Then the blueprint should be used in the following way in the development of such an evaluation instrument.

First, each specific behavior listed on the blueprint is examined to determine whether it is possible to develop a paper-and-pencil test problem which will require the student to perform in the particular way indicated in the chart. If the item cannot be covered in a paper-and-pencil test, a line should be drawn through it to indicate the fact that the test measures only certain aspects of achievement. It is instructive for the teacher to see how many entries remain in the blueprint once it has been decided to appraise student progress in terms of an objective test.

However, care should be exercised in crossing out items from the blueprint since, as will be seen later, a great variety of problem situations can be presented in the form of objective-type test questions. The blueprint shown as an illustration lists few items of behavior which could not be elicited in an objective-test situation. Abilities related to the use of graphs can easily be measured; so too, can the abilities related to the reading of bills, meters, bank accounts and so forth. An example of a behavior tendency that probably could not be measured by an objective test is the item, "Keeps personal relationships frank and friendly."

The weights to be attached to each goal have already been determined in developing the education program. Since it is necessary to enter on the test plan the emphasis or weights to be given to each goal, these weights should be entered on the blueprint beside the goals.

STEP VII. ADDITIONAL FACTORS TO BE SPECIFIED IN THE PLAN FOR EVALUATION

The blueprint alone is not a complete test plan. The complete plan should also include specifications concerning the type or types of measuring techniques that are to be used, samples of the problem situations, specifications concerning the length of the test, the order in which the problems are to be presented, the extent to which the items are to be hard or easy, the procedure to be used for scoring and interpreting the scores, the directions to be given to the student, and any other special conditions which may affect the measuring procedure. Since these factors will be discussed in Chapter VI it is necessary at this point to note only that they should be included in any test plan which must be, as far as possible, a comprehensive specification of all the conditions that are likely to affect measurement.

STEP VIII. THE SELECTION OF MEASURING TECHNIQUES

There is one sound rule to follow in the selection of a measuring technique: the technique selected should be suitable for measuring what it is desired to measure.

During World War II, a directive was issued to all training centers by what was then the Army Service Forces to the effect that all tests given for the purpose of measuring achievement should include completion items, true-false items, and multiple-choice items. This directive was based on the erroneous assumption that all of these forms of tests can measure significant outcomes in all fields. The directive

would have been much more rational if it had stated that instructors should choose the form of test which measured best the desired outcomes of the training mission.

The following chapters are devoted to a discussion of the purposes for which various kinds of objective-test items may be used. Each type of item will be discussed in terms of its usefulness for measuring various outcomes. It is hoped that the material presented will assist the teacher in selecting appropriate evaluation techniques.

WHEN SHOULD THE BLUEPRINT BE PREPARED?

The teacher who has read up to this point may feel that the development of evaluation instruments requires more work than it is feasible to do for evaluation purposes. This would be true if all of the planning here discussed served only the purpose of developing tests. However, that is not the case. Nearly all of it is necessary even when no attempt is made to appraise systematically the outcomes of teaching. Any well-planned course will be based on a series of adequately defined goals and an outline of the content which the teacher feels is most effective in achieving them. The first four steps in the preparation of a test plan should all have been completed as a part of the course plan, for the blueprint of the evaluation procedure is also the blueprint of the course itself.

SUMMARY OF PROCEDURES FOR PLANNING EVALUATION INSTRUMENTS

In this chapter, a procedure for planning an evaluation instrument has been discussed along with certain aspects of

educational philosophy on which the planning procedure is based. In order to help the teacher in the mechanics of planning evaluation instruments, the steps in the procedure are summarized here:

1. The educational goals are summarized in a series of statements. The number of statements is arbitrary, but the author has found eight to fifteen to be a convenient number.

2. Each of the statements listed in the previous step is operationally defined by listing a series of behaviors which characterize the individual in whom the goals have been achieved.[7] These behaviors are referred to as evaluative criteria.

3. Numerical weights are assigned to each goal to indicate the emphasis given to it both in the course and in the over-all evaluation of the extent to which all the goals have been achieved.

4. The content [8] of the course is outlined.

5. A blueprint of the evaluation procedure is prepared. This chart shows the relationship of content to goals and specifies the outcomes that are to be measured.

6. Problem situations are developed.

7. Other factors affecting the evaluation procedure are specified.

8. Appropriate measuring techniques are selected.

[7] Many basic problems in the area of defining goals have not yet been solved or even investigated. For example, there is a need for studying the size of the sample of behaviors that must be specified in order for a domain of behavior to be clearly defined. It is theoretically sound to state that a sufficient number of specific behaviors must be listed in the definition to permit the classification of further behaviors as belonging to or as not belonging to the particular domain. However, the size of the sample which permits this taxonomy needs to be investigated.

[8] The outlining of the experiences through which the goals are to be achieved does not imply that rigid lesson plans have to be followed by the teacher. A well-planned program may allow for flexibility in the classroom.

Chapter Three

Objective-Type Test Questions: Completion and True-False

THE TERM *objective* is used in connection with tests and examinations in a variety of meanings. It refers primarily to a kind of scoring rather than to a kind of test. In an objective test, the rules for scoring are so rigidly and precisely defined that two technicians should be able to score and rescore the same set of examinations and arrive at identical scores. Usually essay examinations are not objective in this sense of the term, though there is no good reason why they should not be built to meet this criterion. Consequently there has been a tendency to use the term *objective* to refer either to the short-answer type of test in which the student's responses are restricted to writing only a few words, or to the completion, the true-false or the best-answer type of test question. In general, in this book the term *objective test* is used in the latter restricted sense, largely because it represents the current usage of the term.

Objective-type test questions will be discussed under three headings: the completion item, the true-false item, and the best-answer item. Almost every type of objective test question is a variation of one of these three forms. In each case,

a discussion will be presented of the kinds of behavior that can be measured with the particular form of item and this will be followed by a discussion of the principles that should be adhered to in the construction of such items.

ORIGIN OF THE OBJECTIVE TEST

The objective-type test was developed largely as an attempt to overcome certain difficulties encountered in the scoring of essay examinations. Investigations have shown again and again that when students are confronted with a broad problem of the type, "Discuss the main causes of World War II," they will vary greatly in the scope of their responses. Some will interpret the problem to mean the factors immediate to the invasion of Poland and the declaration of war by France and Britain. Other students will respond by discussing factors prior to World War I. To such a problem there will be as many different responses as there are students, and how is the examiner to score such variety of products? Unfortunately, what has happened too often is that the examiner has given the best scores to those students who responded in the way the examiner anticipated and the test has thus become a game of guessing what the examiner had in mind when he wrote the problem.

While the example given above is an extreme case, there are few essay questions which do not leave the candidate wondering just what is required. The main consequence of this is that there is a marked element of arbitrariness about the marking of essays, a fact which has been well established by numerous investigations during the past thirty years.

The chief attempt at solving this problem has been to develop essay or free-answer questions which restrict and limit the response of the student. This procedure is discussed

at length in an Appendix. Most essay questions can be broken down effectively into a series of more specific questions, the answers to which can be scored with considerable reliability. The limiting case in such a breakdown is that in which the student has to supply only a single word. In that event the problem is referred to as a completion-test item.

THE COMPLETION ITEM

The completion item is a special case of the free-response item. In general, it does not differ functionally from other types of free-response items except that it is generally used to measure educational outcomes that are psychologically simpler than those that are measured by the essay type of examination.

For a long time the completion-type item has been a stock in trade of the classroom teacher. The reasons for this are obvious, though perhaps unfortunate. There is little doubt that there is no easier method of building examinations than that of sitting down with a textbook, pulling out important sentences, removing key words from those sentences, and then calling the products of this procedure an achievement examination. This method of appraising student progress violates every principle previously discussed and has as its only merit the fact that it enables the teacher to produce, in a very short time, an instrument which is presumed to be a measure of something or other. However, it does not follow that a misused method has no good points to its credit. It does not rule out the possibility that, in certain instances, completion items may measure significant outcomes of education. They should be given consideration as a possible tool in the evaluation procedure.

Common Uses of the Completion Item

1. Completion items are most commonly used for finding out whether the student knows the definitions of terms. Examples of this type of item are given below:

A chemical compound formed by the reaction of an acid and a base is a _____.
(salt)

The technical term for nearsightedness is _____.
(myopia)

The external ear is known as the _____.
(pinna)

Sometimes the completion item calls for a response of several words. The following are examples of this type of item designed to determine whether the student is familiar with certain concepts:

An observation balloon floats in the air because its weight is equal to the weight of the _____.
(air it displaces)
(displaced air)

The volume of a gas at constant pressure is always proportional to its _____.
(absolute temperature)
(temperature on the Kelvin scale)

The energy possessed by a body as a result of its motion is called _____.
(kinetic energy)

2. In another form of the completion item, the subject is presented with a passage from which a number of key words have been deleted. Usually in such cases, the examiner has attempted to leave enough of the essential material in the paragraph so that the person who has an adequate background can easily fill in the missing key words.[1] The following are examples of this type of completion-test question, which is commonly called the *connected-discourse* type:

Atrophy of the _____ gland in childhood produces a
(thyroid)
condition known as cretinism. A similar atrophy in the adult produces _____.
(myxedema)

From time to time the gene will change its structure, or one or more adjoining genes will become detached from their _____ and be lost. When such a change
(chromosomes)
takes place, it is called a _____.
(mutation)

The quantity of matter in a body is referred to as its _____, while the force exerted on the body by gravity
(mass)
is its _____.
(weight)

It will be noted from these examples and from general observation of the completion-type item that one of its main uses is for measuring active vocabulary, that is to say vocabulary which can be recalled when the situation re-

[1] A major advantage of the connected-discourse type over the isolated-sentence type is that it provides the examinee with more adequate cues. Isolated sentences with a single word omitted in each often become tests of reasoning, because they do not provide enough information for the examinee to fill in the blanks on the basis of knowledge of terms alone.

quires the use of a given term. Probably the main goals that the completion-type item is most successful in measuring are those related to the acquisition and use of vocabulary.

3. The connected-discourse type of completion item is a poor measure of vocabulary when it introduces a reasoning factor which cannot be adequately controlled. The following problem, which appeared in a test of American History, illustrates how a completion item may introduce a reasoning factor which is largely irrelevant to the goals of most history courses. This does not mean that thinking skills are not important outcomes of the teaching of history, for they may be. However, it does mean that there is little point in making the student arrive at the answer to a question by a long deductive process which has nothing to do with the outcome of the teaching of history.

> By unanimous consent _____ became the leader of the Republicans. After some study of the fiscal system developed by _____, he became an opponent of all the leading measures fostered by the Secretary of the Treasury, including the establishment of the Bank.

In order to fill in the blanks, it is necessary for the student to deduce that the paragraph refers to the early nineteenth century. He can do this only by noting that the Bank referred to is probably the United States Bank, founded by a bill passed by the Senate in 1791. The student must then go on to deduce that the first blank space is for the name of a President and that he must think of a President who was a friend of the farmer and opposed to commercial interests. Note also that he can find the name of the Secretary of the Treasury only if he has correctly deduced the period of history discussed. In this completion test, the reasoning involved has a strong element of triviality about it.

However, connected-discourse completion items can be

used appropriately in tests designed to measure deductive-reasoning ability, and they have been used successfully in a number of so-called tests of intelligence. But other techniques are usually better for measuring thinking skills related to subject matter. The student of history may be expected to develop the ability to make important and significant deductions from a given body of historical data, but it is not reasonable to expect him to make the kind of academic and useless deductions which he must make to solve the problem cited above.

Similarly, if the reader studies the following series of completion items he will see that the kind of reasoning necessary for solving the problem is largely irrelevant to the commonly accepted goals of the teaching of American History.

> With characteristic abruptness, _____ decided to sell to the United States every inch of the territory so recently wrung from _____. When the news crossed the Atlantic, no one was more astounded than _____. At first he prepared an amendment to the Constitution which would authorize the purchase, but later decided that this procedure would be impractical.[2]

If additional cues are provided, the entire function of the item may change. Compare in this respect the following version of the completion item under discussion with that given above:

> With characteristic abruptness, Napoleon decided to sell to the United States every inch of the territory so recently wrung from _____. When the news crossed the Atlantic, no one was more astounded than President _____.

[2] The three blanks can be properly filled with the names Napoleon, Spain, and Jefferson.

The latter version is not a measure of thinking skill; it has become a measure of information.

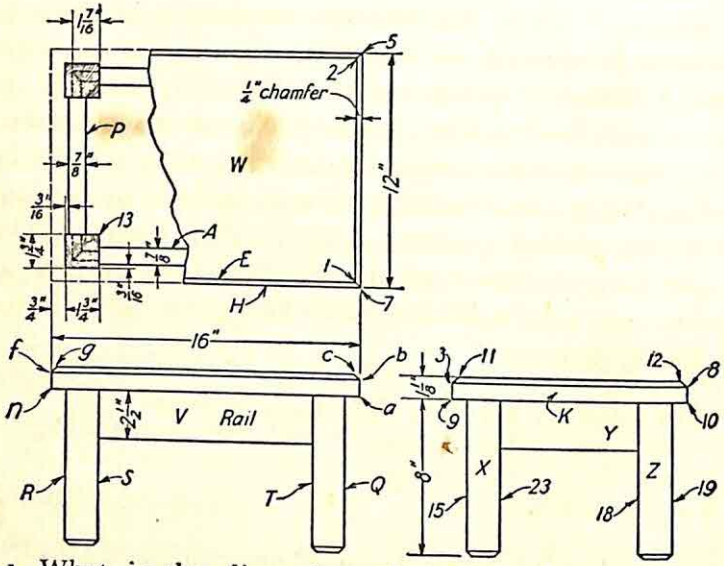
4. One of the few legitimate uses of completion items for appraising thinking skills related to subject matter is in the measurement of the outcomes of teaching elementary mathematics, general mathematics, and subjects requiring mathematical solutions to problems. Probably the strongest argument in favor of the completion item in such cases is that it permits the student to manifest precisely those behaviors which are acceptable as evidence of the achievement of the desired outcomes. An item of the following type measures directly the student's ability to perform one important task of an intelligent consumer:

> Eggs weighing 1½ oz. each sell at $0.50 per dozen while eggs weighing 2 oz. each sell at $0.64 per dozen. What is the difference in the price per pound of the two grades of eggs?
>
> _____ cents

Presented in this form, the problem requires the student to go through the entire process which an intelligent consumer must complete in determining which grade of eggs is the best value for the money. The student cannot arrive at the solution by an indirect method, as he might if the problem were presented in a multiple-choice form. However, as will be seen later, there are certain difficulties involved in the scoring of such problems.

The following problem illustrates how a whole series of completion problems may be based on a single major problem. The items are not stated in the incomplete-statement form, but in question form. In this case the form of the response is the same regardless of the method of presentation used. Each question could be converted into an incomplete statement.

Answer the questions concerning specifications for a stool as shown in the following drawing. Certain points and edges are indicated by letters and numbers on the drawing.[3]



1. What is the dimension of the chamfer shown in the front view? _____ in.
2. How much does the top of the stool project beyond the front leg? _____ in.
3. What point on the top view is represented by line S in the front view? _____
4. What is the distance between the inside of the legs in the side view? _____ in.
5. What line in the top view represents the back surface of rail V? _____
6. What point in the top view represents point 10 shown in the side view? _____

[3] This drawing was made and the sample questions were written by J. M. Gray of the Brooklyn Technical High School. In William A. Brownell *et al., The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, p. 292). The University of Chicago Press, 1946.

This problem included a long series of additional questions which do not have to be reproduced for illustrative purposes.

5. Similarly, completion problems are commonly and legitimately used by mathematics teachers to measure the student's facility with the mechanics of manipulating mathematical symbols. Although there are teachers who feel that facility with the mechanical manipulation of symbols may be an important outcome, the achievement of this outcome does not imply that the student has acquired understanding. Consequently, completion items of the following type should be considered to measure outcomes which are stressed only by a limited number of teachers in a limited number of situations:

The coefficient of a^4b^3 in the expression $(a + b)^7$ is _____ _____.

If $y = x^2\sqrt{5 - 2x}$, then $\frac{dy}{dx} =$ _____.

The factors of the expression $9x^2 - 3x - 2$ are (_____) and (_____).

Items of this type have been commonly used for measuring the outcomes of college courses in mathematics. It should be remembered that they are generally used with the assumption that the ability to manipulate the mathematical symbols is evidence of understanding. This assumption is rarely justifiable.

6. Items similar to those used by mathematics teachers have been used for measuring the outcomes of a first course in chemistry. In this type of item the student is required to complete a chemical equation when only one side of the equation is given:

$NaPO_3 + CO \rightarrow$ _____ + _____

$PCl_3 + 3H_2O \rightarrow$ _____ + _____

Weaknesses of Completion-Type Items

The discussion up to this point presents the various common uses of completion items and the kinds of outcomes they attempt to measure. There are, however, certain major defects in completion items which limit them in value. These defects are discussed below:

1. It is almost impossible to develop a completion item for which the correct answer has only one possible form. For example, to the simple problem in which students are asked to find the cost to the puchaser of 15 oranges at 40 cents per dozen, a whole series of different forms of the correct answer will be given. Students will give several unquestionably acceptable answers, such as "50¢," "fifty cents," "$0.50," and "½ dollar." Then there will be students who will give answers such as "fifty," "50," and "0.50." In addition, there will be those of limited literacy and those who are a little facetious who will give answers such as "fifty sense," "four bits," "half a buck," and other colloquial forms which convey the right concept even if they do not convey it in an academic form. Usually it is almost impossible to predict the variety of answers that will be provided by a group that takes a completion-test item. In the simple arithmetic problem described above, a group of 200 students may give as many as 30 different answers, which vary in their degree of acceptability.

Similar difficulties are found in the scoring of almost all types of completion items, but the poorer the item the more such difficulties multiply. The result of this deficiency in completion items is that the scoring process often involves a strong subjective element.

2. Since it is usual in the connected-discourse type of completion item to omit key words, it is often difficult to supply the subject with adequate cues. If adequate cues are given.

the answer to the item becomes obvious or the item acquires a trivial character. On the other hand, when inadequate cues are provided, the item becomes largely a test of reasoning and is unsuitable for measuring any outcome of instruction in almost any field.

3. The completion problem is usually very unrealistic. That is to say, it is only remotely related to the behaviors that are acceptable as evidences of achievement. There are relatively few occasions in daily life when it is necessary to fill in the gaps in discourse or to supply the last word of a statement. In most life situations for which the school provides preparation, the student is faced with a problem and sees several possible solutions from which he must choose the one which will work best in practice. There are very few completion tests of the type which omit a single word which do not give the appearance of extreme triviality. This comment does not apply to the essay examination. As will be seen later, the essay examination may provide a series of thoroughly realistic situations of a kind which a great many educational programs prepare people to meet.

Rules to Follow in Constructing Completion-Test Questions

The preceding discussion indicates that there are not many situations of which it can be claimed that the completion test provides a better measure of a significant outcome of instruction than can be obtained by other techniques. However, there are a few situations in which the completion item can be used appropriately, and in such cases these rules should be followed:

1. Keep the ratio of words given to words omitted very high because, if too many words are omitted, the meaning of the whole will be obscure.

2. Usually avoid taking statements directly from textbooks, and never take statements from textbooks which the students have used.

3. The blanks should refer only to omitted key words. Never leave blanks for such words as *a, an,* or *the.*

4. If the answer involves numerical units—e.g., 20 yards, 72 miles, 8 pints—do not leave blanks for the names of the units. It is usually wise to designate the names of the units in terms of which the answers are to be given. Alternatively, the numerical answers may be given and the student may be asked to fill in the names of the units.

5. Remember that unless a completion item is very carefully constructed it is likely to become a measure of general reasoning ability. Remember, also, that many well-known intelligence tests include completion items for measuring reasoning ability.[4]

THE TRUE-FALSE ITEM

For many teachers the true-false item has been the main instrument of objective measurement, and many of those who have had little use for objective tests have based their attitude on their experience with true-false tests alone. This appraisal of all so-called objective tests is unsound, because experience with true-false tests indicates that, of all techniques of objective measurement, the true-false item is the least flexible, the most limited in the behaviors it can measure, and one of the most difficult to handle effectively. It should be noted in this connection that the least useful form of the objective-test question is the one that is most commonly used by teachers.

[4] The *C.A.V.D. Test* is an example of a test which uses completion items for measuring reasoning ability.

Forms and Uses of the True-False Item

1. In the simplest form of true-false item a simple statement is presented, which a student must appraise as being either true or false. The following illustrations are quite characteristic of items appearing in examinations; they are given here without any evaluation of their merits.

Thyroxin has a function related to calcium metabolism.

One function of vitamin B complex is to prevent scurvy.

The Reconstruction Finance Corporation was first developed under the administration of Herbert Hoover.

Through taxes communities purchase services which each member could purchase individually for himself.

The items above test the student's knowledge or ignorance of some established facts.

2. A more complex variation of the preceding type of item occurs when the student is presented *not* with a simple problem of the type "X is Y," or "X is used for making Y," or "X prevents Y," but with a proposition of the type "X is Y because X is also Z." Examples of this type of item are given below:

Citrus fruits are desirable parts of the diet because they help prevent scurvy.

If an electric motor stalls it may burn out, because stalling reduces the back emf in the windings.

Electric light bulbs burn out because the filament slowly becomes oxidized.

One important reason for the adoption of the European Recovery Program was the fear which the United States felt of communism.

In items of the kind given above it is usual for the first part of the item to be true and for the reason given in the second part of the item to be either correct or incorrect. In such cases the item may measure the student's insight into the reasons underlying a phenomena, and the item *may* measure something more than memory for facts. Items of this kind are likely to be confusing and time-consuming, since the student must determine not only whether each of the propositions involved is true but also whether the relationship between the propositions is true. In order to simplify this situation it is often desirable to tell the student that both of the propositions are true and that he has only to examine the relationship between them. In any case special care is required in the preparation of this type of item since it is liable to become a measure of reading comprehension and of nothing else.

3. Students are sometimes asked to classify a statement as true, false, or, if it cannot be properly classified as either true or false, as "doubtful." Examine, for example, the following pair of statements:

> Franklin D. Roosevelt was elected four times as President of the United States.
>
> Franklin D. Roosevelt was the greatest of all Presidents of the United States.

The first of the two statements can be easily classified as true or false, but the second statement is a value judgment and is based on opinion. There is no objective basis for determining whether it is true or false and consequently it should be classified in the "doubtful" category. Items of this kind are of use mainly for determining whether a student can discriminate between value judgments and statements of fact.

4. Occasionally there is a further elaboration of the student's response to a true-false item, and he is asked to classify the item according to whether it is true, probably true, probably false, or false.

5. Sometimes the student is asked not only to classify a statement as true or false but also to classify in the same way the converse of the statement. Items of this kind have appeared in tests of logic, mathematics, and philosophy. For example:

All men are bipeds. (The converse would be "All bipeds are men.")

No amoebae are parasites. (The converse would be "No parasites are amoebae.")

The present writer feels that in most tests in which this form of item has been used the outcomes measured are rather trivial.

6. In some tests a student is given a statement and asked to cross out, or in some cases to change, the one word in it that makes it false, or to select from the words which are underlined the one that makes it false. Examples of this type of item are given below:

Bile$_{1}$, the secretion of the liver$_{2}$, contains no enzyme$_{3}$.

The stomach$_{1}$ is connected at its posterior end with the large intestine$_{2}$ which in the frog is subdivided into two portions, the rectum$_{3}$ and the cloaca$_{4}$.

It should be noted that, in order for items of this kind to be effective, it is usually necessary for them to be long and

to contain several technical terms. Consider, for example, the following item:

The largest division of the human brain is the cerebellum.

In this item there is only one word that is defined either correctly or incorrectly, namely the word *cerebellum*. Consequently, the student may as well be asked to judge the statement as true or false. Also, more than one word may be changed to make this statement true. If *cerebellum* were changed to *cerebrum*, or if *human* were changed to *fish*, the statement would become true.

7. Sometimes a true-false test is elaborated so that several true-false statements are based on a relatively large quantity of data. For example, an experiment in biology may be described, and a series of statements may be given each of which represents a possible conclusion drawn from the experiment. The student is required to evaluate each of these conclusions and to determine whether it can legitimately be drawn from the data. In another variation of this form of item, the student is given some data and a conclusion drawn from it. He is then required to identify from a list of possible assumptions those that have to be made in drawing the given conclusion from the given data. It should be noted that these test situations use derivatives of the true-false form of item rather than the simple true-false item form. In these situations the student does not strictly make a judgment of truth or falsity, he judges whether a conclusion is sound or an assumption is necessary. The outcomes measured by such items are usually considered to be aspects of critical thinking and are significant outcomes in a great variety of courses. Probably they are among the most significant outcomes that the true-false test can measure. Some examples of measuring instruments of this kind are given below:

Example I.[5] The italicized statement at the end of the problem is assumed to be a correct answer. You are to *explain* the italicized conclusion by selecting statements from the list following the problem. [The student checks the explanations.]

If a person is planning to bathe in the sun, at what time of the day is he most likely to receive a severe sunburn? *He is most likely to receive a severe sunburn in the middle of the day* (11 A.M. to 1 P.M.) *because:*

( ) We are slightly closer to the sun at noon than in the morning or afternoon.

( ) The noon sun will produce more "burn" than the morning or afternoon sun.

( ) When the sun's rays fall directly (straight down) on a surface, more energy is received by that surface than when the sun's rays fall obliquely on the surface.

( ) When the sun's ray fall directly (straight down) on a surface, less sunshine is reflected from the surface than when the sun's rays fall obliquely on that surface.

( ) When the sun is directly overhead the sun's rays pass through less absorbing atmosphere than when the sun is lower in the sky.

( ) Just as a bullet shot straight into a block of wood penetrates farther into the wood, so will the direct rays at noon penetrate more deeply into the skin.

( ) The air is usually warmer at noon than at other times of the day.

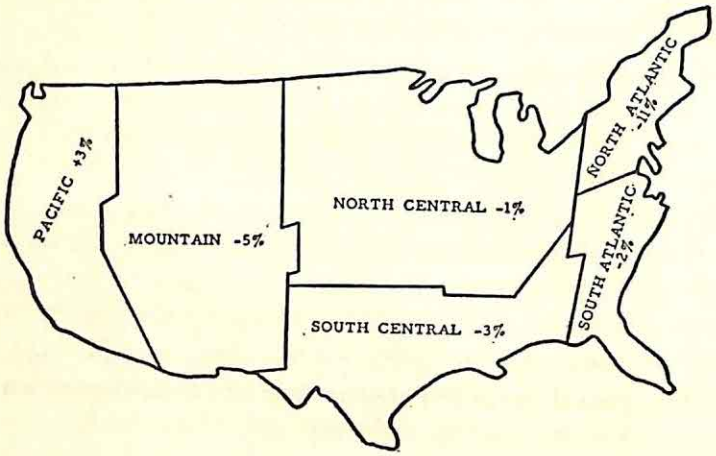
( ) The ultraviolet of the sunlight is mainly responsible for sunburn.

[5] In William A. Brownell *et al., The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, pp. 114–115). The University of Chicago Press, 1946.

Example II.[6] In this example, the student classifies an interpretation of data in one of three ways, such as the following:

A—enough information is given to make the statement *true*
U—not enough information is given to decide
D—enough information is given to make the statement *false*

PER CENT OF CHANGE IN MOTOR FATALITIES (DEATHS) IN THE REGIONS OF THE UNITED STATES IN 1939 AS COMPARED WITH 1938



The map above shows for each region the percentage of increase or decrease in motor fatalities in 1939 as compared with 1938. A minus per cent means a decrease in motor fatalities; a plus per cent means an increase in motor fatalities. The per cent for the United States as a whole for 1939 is — 2%, which means there were 2% less

[6] In William A. Brownell *et al.*, *The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, p. 123). The University of Chicago Press, 1946.

motor fatalities in the whole of the United States in 1939 than in 1938.

Statements

1. In 1939 there was a decrease in motor fatalities in most of the regions of the United States.
2. The per cent of decrease in motor fatalities was greater in the South Central region than in the North Central region.
3. The general decrease in motor-death toll in the United States in 1939 is due in part to improved traffic laws.
4. The largest decrease in the per cent of motor fatalities in the regions of the United States occurred in the North Atlantic region.
5. In 1939, there were three times as many deaths due to motor accidents in the South Central region as in the North Central region.
6. There was a two per cent decrease in motor fatalities in Florida.
7. There were more motor fatalities in the Pacific region in 1939 than there were in 1938.
8. In 1939, more people were killed in motor accidents in the Pacific region than in the North Atlantic region.
9. The death rate due to motor fatalities is decreasing from year to year.
10. Such records will encourage traveling in the North Atlantic region.

Limitations and Weaknesses of the True-False Item

True-false items have limited worth as evaluation instruments for the following reasons:

1. The true-false item usually fails to present a realistic type of problem, that is to say, a problem similar to those which the educational program has taught the student to solve. Rarely in life is a person faced with the problem of deciding whether a statement is true or false.

2. The true-false items tends to be what has been called "textbookish," that is to say, it tends to measure the extent to which the student remembers a particular textbook he has read. True-false items are not necessarily "textbookish," but they frequently are because of the techniques adopted by teachers in preparing such tests. A true-false test based on textbook statements is simply one which requires the student to discriminate between his text and variation of it. In such an examination the student who has carefully memorized the text, maybe without digesting it, often has considerable advantage over the student who has digested his text without any deliberate effort at memorizing it.

3. The true-false item is limited in the outcomes it can measure. Because of this limitation, most tests consisting only of true-false items measure only a few of the outcomes which it would be desirable to measure. Usually, such tests do not measure some of the more important outcomes but are limited to outcomes related to the acquisition of the terms and concepts in the field.

4. True-false items provide rather low reliability per item.[7] This means that it is necessary to include a very large number of true-false items in a test in order to produce

[7] *The Meaning of Reliability and Validity.* All educational measurement is contaminated with error and one cannot be sure that if a test could be repeated under apparently the same conditions it would yield the same scores. The correlation between the scores obtained on different administrations can be estimated, and this estimated correlation is referred to as the coefficient of reliability. However, even when a test can be expected to provide highly consistent scores it may still fail to measure what it is supposed to measure, that is to say it may not be valid. A reliable test is not necessarily valid, but a valid test is necessarily reliable.

reasonable reliability. Of course, it must be remembered that even if the true-false test can be made reliable, its validity and comprehensiveness may often be questioned. In this connection, it may be noted that the type of true-false test commonly made by teachers consists of only 50 to 75 items and is, in most cases, practically worthless as a measuring instrument.

5. Contrary to common belief, good true-false items which measure significant outcomes are particularly hard to write. If the individual who proposes to use true-false items goes about the process in a systematic manner and avoids the common practice of simply pulling statements from textbooks, he will find that it is extremely difficult to build items which attempt to measure the kinds of outcomes he believes to be important.

Rules for Preparing True-False Items

Although the true-false type of item is not generally recommended, there are cases where it can be used appropriately. Under such conditions true-false items should be built on the basis of the following rules:

1. Avoid statements that are broad generalizations, since such statements are practically always false. Statements such as "All physical qualities are inherited" or "The sole cause of rickets is a deficiency of Vitamin D" are necessarily false and are given away by the use of the words *all* and *sole*. General statements that include such words as *never, always, sole,* and other absolute or all inclusive terms should be avoided, since many students will recognize them as false.

In the same way that *always, never,* and similar terms may indicate that a statement is false, so too will *generally, sometimes, usually, most,* and similar words indicate that a statement is probably true. The best practice is probably to use

the latter variety of modifier in both true and false statements, but occasionally a universal such as *all* or *always* may be introduced into a true statement in order that the student who has no basis for responding adequately may be led into believing that it is false.

2. ***Avoid trivial and meaningless statements.*** This rule may seem too obvious to be stated, but it is the one that is most frequently infringed in tests. Meaningless statements are commonly included and keyed as false statements, for it is easy to divest an elaborately structured statement of meaning by changing some of the words in it and to use the result as a false statement. The following are typical examples of relatively trivial or meaningless statements from true-false tests:

> Well-written material expresses ideas in the form most appropriate to the concept which the author has in mind.
>
> All behaviors are necessarily neural in origin because activity is intrinsic in the nature of the arc mechanisms.
>
> Unless a novel has a basic emotional appeal it will not be enjoyed.

Now consider the following item, which is an example of an obviously true statement. It appeared in a college examination.

> A student can improve the efficiency with which he learns by practicing good study habits.

The student who reads the item above with understanding will inevitably recognize its truth. Yet it was intended to be used to measure the student's knowledge of good study habits. The teacher who wrote this item failed to make an

analysis of the important understandings which she was trying to develop in her students.

Trivial statements are also frequently encountered in true-false tests. The present writer has actually encountered in a true-false test the following statement:

The number of pages in your textbook is 465.

Similarly, the statement,

Isaac Newton lived from 1642 to 1728,

is completely trivial as a false statement. Apart from the fact that the exact dates of birth and death of Newton may be unimportant, it is ridiculous to attempt to trip the student on the fact that Newton died in 1727 and not in 1728.

There are frequent occurrences of statements equally trivial but more closely related to the actual subject matter. Sometimes items are relatively trivial because they cover the least rather than the most significant facts in a given situation. The following item illustrates this kind of triviality:

The first person to speak in Oscar Wilde's play, *The Importance of Being Earnest,* is Ernest himself.

The item covers a point about the play which is completely inconsequential. There are many things about the play which are more important than who opens the dialogue, and the above question could not be justified in any test.

3. In general, negative statements should be avoided in true-false items, chiefly because they are often misread. Even with excellent readers there seems to be a marked tendency for negatives to be overlooked in short statements. Some test writers try to overcome this reading difficulty by

setting all negative words, such as *not,* in capital letters or by underlining them. This practice does partially eliminate the difficulty, but certain terms in which the negative is included in the first syllable also provide major reading difficulties. An example of the latter is the word *undesirable* which is likely to be read as *desirable.*

4. Avoid developing true-false tests by extracting statements from textbooks. This point has been stressed but deserves to be stressed again. It is a very common but undesirable practice to develop true-false tests by the simple and expedient procedure of first lifting from a textbook all statements which are meaningful out of context, and then reversing the meaning of half of these statements and assembling the collection as a test. This practice has the unfortunate result of emphasizing as a major objective the memorization of material, and the test may be a good measure of that particular outcome. While there are a few fields in which it is customary to stress memory of facts as a major outcome, in most areas of teaching it is unfortunate to stress sheer memory of content.

5. Avoid statements that are partly true and partly false. Such statements are very disturbing to the student and produce an undesirable emotional response. Statements of the following type illustrate this unwise practice:

> One of the founders of the Soviet Republic was Lenin and he is living in Moscow at the present time.
>
> Most of the radiation from the sun is invisible and fails to penetrate the atmosphere.

Statements such as the above are much more effective as measuring instruments if they are divided into two parts and if the student is asked to judge the truth or falsity of each part.

6. Avoid ambiguous statements, since they may be either true or false according to the particular interpretation given to them. It should be noted that a test writer is a very poor judge of the clarity or ambiguity of the test problems which he himself has written. Ambiguous statements can be detected most easily by having the test material reviewed by others and by observing the interpretations they give to each statement. Item analyses, which are discussed later, also help in the identification of ambiguous items. The following true-false statement contains an ambiguity which the writer of the item did not identify:

> The cost of a house depends upon the time taken to repay the mortgage loan.

Presumably, the author of this item had in mind the fact that a long-term mortgage adds to the total amount a person pays for a house. However, some may interpret this statement as meaning that the cost of building a house is related to the term of the mortgage. If the word *cost* refers to the building cost, then the statement is false. Consequently, if this item is to be used, some word other than *cost* should be employed. Other examples of common types of ambiguous statements are given below:

> There is a general tendency for human beings to increase in size.

> The cost of 10 pairs of shoes selling at $8.00 per pair is $80.00.

At least some of the ambiguity provided by the true-false test situation can be avoided by directing the student to mark a statement as true only if it is completely true under all circumstances. This helps to remove some of the inse-

curity which students feel when they encounter an item which is true except under one unusual condition.

7. Avoid including two or more ideas in one statement. A special case of this kind has already been considered, in which items included both propositions that are true and propositions that are false. However, it is a common but undesirable practice to include several true facts or several false propositions in one statement. Examples of such items are given below:

> Lenin, Trotsky, and Stalin were all leaders in the Russian revolution.
>
> Radio-active isotopes have been used both in the diagnosis and in the treatment of cancer.

If each of these statements were divided up into a series of propositions the measuring procedure would be improved. Statements that include more than one idea are most justified in the type of items already considered, in which the student is asked to judge the correctness of a reason given for a particular phenomenon. A more complicated example of the multiple-idea defect is given below. If the item were broken down into a series of separate items, it would be possible to determine which factors in the item were understood by the student and which were not.

> Oil occurs not only on simple anticlines but also on many other types of structural features including terraces, anticlinal noses, faults, unconformities, salt domes, lenticular sands, and buried hills.

The following item is somewhat simpler than the one given above, but it should nevertheless be broken down into two independent items:

In absorption spectrophotometry, light waves of every length in the desired region are sent through the semi-transparent material to be analyzed and the spectrograph is used to determine how much of the light of each wave length has been absorbed by the material.

8. If items express opinions, it is important to attribute the opinions to some source. For example, it is quite unfair to present the student with the statement:

Shakespeare is the greatest dramatist that the world has so far produced.

It is not possible to ask the student whether this statement is true or false, since opinions could differ widely on the particular point covered by the item. On the other hand, it might be reasonable to present the student with the statement:

Golancz considers Shakespeare to be the greatest dramatist that the world has yet produced.

Similarly, in all items that present value judgments, the value judgments should be attributed to a definite authority. This is particularly important in measuring outcomes in fields such as literature and art and in every area in which judgments are made in terms other than those of verifiable principles.

9. Long and involved statements should never appear in true-false tests, because they tend to measure certain aspects of reading comprehension that are much better measured by other techniques. Read, for example, the following statement and note that the difficulty in understanding it is mainly a result of the complex verbiage, for the ideas are fairly simple. The appraisal of their truth or

falsity requires careful analysis of the complex sentence structure.

> Overlearning pays because there is evidence to show that material 100 per cent overlearned is retained much more than 100 per cent better than material learned to the point where the correct word of a series is anticipated in 100 per cent of the cases.

This tongue-twister requires the student to disentangle the ideas before he can determine that the statement is rather meaningless. If the teacher was trying to discover whether the student knew that overlearning pays, it might have been better for him to ask the student to evaluate such statements as, "In most learning situations, overlearning is a worthwhile investment of time" or "In learning material in school, overlearning is a poor investment of time."

10. Trick questions and catch questions should never have a place in an examination. Consider the statement:

> A man in the Northern hemisphere walks ten miles due north, ten miles due east, ten miles due south, ten miles due west, and arrives back where he started from.

If this statement were given as part of a geometry examination, it would be answered correctly only by those who were aware of the catch in the problem. The statement is false, because the man would not arrive back at the point from which he started but would finish at a position east of that point, if he were in the Northern hemisphere, since the surface of the earth is curved. High-school students who did not take account of this particular fact would be misled into thinking that the statement was true and few would suspect any catch in the problem. To most of them it would seem that the test problem was unnecessarily easy. Catch prob-

lems are occasionally used to measure thinking skills, but in actual practice they are poor measures of those skills because they usually require a greater ability to reason than the student can be expected to show. Consequently they tend to measure only whether or not the student is familiar with the particular idiosyncrasy involved.

11. True Items and False Items should have the Same Average Length. Many teacher-made tests of the true-false type tend to give away the right answers because the true items are somewhat longer than the false items. The reason for this is that short and concise false statements that are unquestionably false are easily written, but it is more difficult to write true items briefly and concisely. In order to make a statement indubitably true it is usually necessary to expand it and to use qualifying terms and clauses.

In order to remedy this common deficiency, it is necessary as a rule to expand some of the false statements in order to make them approximately equal to the true statements in average length.

THE VALUE OF COMPLETION AND TRUE-FALSE TESTS TO THE TEACHER

So far as it is possible to make an over-all appraisal of completion and true-false test items it may be said that they have limited value as measuring devices. This is well reflected in the fact that they have disappeared almost entirely from published tests of achievement in subject-matter fields, because most test writers believe that nearly all the outcomes which they attempt to measure can be measured better by other forms of tests, and particularly by the type of test discussed in the next chapter.

Chapter Four

Objective-Type Test Questions: Best-Answer or Multiple-Choice

IN RECENT YEARS, published achievement tests have tended to use multiple-choice or best-answer types of test questions almost to the exclusion of all others. The reason is that these test questions have certain advantages not possessed by other types. Nearly every useful function which the true-false item or the completion item can perform is better executed by multiple-choice questions, although the latter type of test question does have certain definite limitations which will be considered later.

TERMINOLOGY

The multiple-choice test question consists of two parts. First, it presents a statement of a problem, called the *lead* or *stem* of the item. Sometimes the lead is stated in a question form. Sometimes it is stated in an incomplete-sentence form.[1] The following items illustrate the question type and the incomplete-statement type of lead:

[1] For reasons which will become apparent at a later point the type of lead which consists of a declarative statement is not generally acceptable.

Which one of the following types of test is generally considered to be *least* valid?

a. Intelligence tests
b. Tests of mechanical ability
c. Personality tests
d. Tests of sensory-motor coordination

The first step in the building of a trade test should be to

a. make an analysis of the job by observing men on the job.
b. survey training programs.
c. develop some test problems for experimental use.
d. examine books on the trade.

The second part of the multiple-choice item consists of a series of suggested answers (as in the examples above) of which usually only one is correct. These suggested answers are sometimes referred to as *alternatives,* and the incorrect alternatives are sometimes called *decoys* or *distractors.*

It should be noted that the multiple-choice type of item is now generally considered to be a combination of a problem and a series of suggested solutions. For reasons which will be considered later, it is usually pedagogically unsound to develop multiple-choice questions each of which consists merely of an incomplete statement, meaningless in itself, for which various suggested endings are supplied and from which the student must select the best ending. This approach to multiple-choice questions is meaningless because the goals of education are not concerned with improving the ability of the student to find the best endings for statements. However, the goals of education are very largely those of developing in students the ability to solve problems and to recognize correct solutions.

MERITS OF THE MULTIPLE-CHOICE QUESTION

The multiple-choice test question has several merits that have made it a major device in the appraisal of student achievement. These advantages include the following:

1. Since multiple-choice questions present the student with definite problems and since teaching is at least partly concerned with problem solving, they do bear some relationship to the desired outcomes of instruction. They provide a much more realistic situation than true-false items, which are restricted to measuring the student's ability to judge a statement as being true or false.

2. The multiple-choice type of problem presents a very flexible kind of problem situation and, contrary to a common misconception, it can be used to appraise thinking skills as well as simple recognition skills. There is nothing unrealistic about the way in which the student responds, though many critics feel that the free-answer test represents something much nearer to the situations that arise in daily life than does the multiple-choice type of test. However, there are two sides to this question. In most problems that are commonly encountered in life, the *possible* solutions are evident and the problem is largely that of selecting the right solution. Most of the mistakes people make in life are not a result of failure to consider the correct solution to a problem as a possible one; they result more frequently from failure to recognize the correct solution as the best and the resulting choice of an inferior alternative. Very few situations are ever encountered, outside of scientific work, in which the individual does not have to make a choice from the alternatives which present themselves. Consequently, the multiple-choice test problem is not so artificial as it may

seem to be at first sight. In most multiple-choice problems in which the student has to weigh the relative merits of the various solutions, the tasks he performs are not very different from those he must undertake in daily life.

3. The multiple-choice item which has four or more alternatives usually provides greater test reliability per item than the true-false type of test. In general, it is easier to achieve a reliable test if multiple-choice items are used than if other types of items are used.

4. Contrary to the belief of many, it is generally considered much easier to develop valid multiple-choice test questions than it is to develop valid questions of almost any other type. As compared with the multiple-choice item, the true-false item is much harder to handle, since the test writer usually finds it difficult to control the things that are to be measured.

WEAKNESSES OF THE MULTIPLE-CHOICE ITEM AS A MEASURING DEVICE

It is somewhat unfortunate that many teachers who commonly use multiple-choice questions for measuring achievement fail to recognize the following limitations of this form of measurement:

1. It is fair to say that, for the most part, objective examinations represent substitutes for better forms of measurement. For example, it is obvious that, if a high-school program is designed to improve the health habits of the student, a health problem in an objective test often produces fundamentally different responses than it does when it is presented in an out-of-school situation. The student who is able to give the correct solutions to paper-and-pencil problems concerning nutrition and the maintenance of health

may nevertheless be the student who does least to protect his own well-being. The student who knows the advantages of a good breakfast is not necessarily the one who gets up early enough to eat the proper meal. The student who knows how to purchase wisely may fritter away his money on worthless things. There is often a discrepancy between a paper-and-pencil response and a response outside of the classroom. While the multiple-choice test question is the most flexible of all objective-test forms, it shares this disadvantage with other forms of paper-and-pencil items. There is no guarantee that the choices made in the test situation will actually correspond to choices made at other times.

2. While a problem presented in written form may apparently resemble the problem which an educational program has attempted to teach a student to face, it may omit certain essential elements. For example, the following problem was given to a group of supervisors of clerks:

> Two of your clerks are constantly quarreling over whether one or the other is responsible for a particular job. The result is that many jobs which you need to get out are delayed unnecessarily. What should you do in this situation?

In the paper-and-pencil situation, this problem may be calmly considered and the best solution may be selected without any emotional tensions interfering with the reasoning process. However, if this situation were faced by a supervisor, it is quite possible that the frustrations and irritations resulting from the situation might be such as to provoke an outburst of wrath rather than a rational solution to the problem. It is to be noted that the paper-and-pencil situation excludes certain elements which might in practice inhibit or facilitate the appearance of an acceptable response. Paper-and-pencil test questions on how to handle emergency

situations are not likely to provide satisfactory evidence of the ability to handle such situations in real life. The paper-and-pencil problem does not include the essential element of surprise, which may result in a disorganized response.[2]

3. Multiple-choice questions do not seem to lend themselves to the measurement of creative abilities. In measuring the capability of the student to do creative work, it is necessary to present him with a problem of creating either ideas or objects which will perform certain functions. The process of creation is fundamentally different from that of selecting from a set of solutions the one that applies best to the particular problem. At the present time it seems that creative abilities can be measured only through a free-response type of problem situation in which only the problem places any restriction on the performance of the individual. Consequently, the traditional type of essay examination and performance tests of various kinds offer the greatest promise of measuring the student's creative abilities. While the criticism that multiple-choice tests do not measure creative abilities is fundamentally sound, it is often erroneously taken to mean that objective tests should be discarded. It must be recognized that there are relatively few courses below the junior or senior year in college in which originality and creativity can be considered as major outcomes. Certainly, most courses in mathematics, social studies, foreign languages, and science do not attempt to stimulate original thought, and the measurement of creativity is quite inappropriate in any appraisal procedure in these fields. In the language arts, on the other hand, it is usually considered that the student may be expected to be creative at a very early age and consequently, in that field, it is appropriate to measure creativeness as a major outcome. How-

[2] Note also that it seldom includes the extraneous, irrelevant information which may be a potent distractor in a life situation.

ever, the limitation on the use of multiple-choice problems under discussion is not a serious one in appraising the outcomes of most fields.

4. The multiple-choice type of question has limited value for measuring the ability of students to organize their ideas. Various types of multiple-choice items have been developed for measuring this outcome but they usually involve such indirect procedures that their validity may be questioned. There seems to be little doubt that the ability to organize ideas in almost any field can be acquired to some extent through adequate training. In general, it seems desirable to avoid using the multiple-choice type of question for measuring this outcome, since there are better ways in which it can be measured.

FORMS AND USES OF MULTIPLE-CHOICE QUESTIONS

The forms in which multiple-choice questions are stated are so varied that it would require a volume to provide illustrations of all of them. The examples given here serve to illustrate some of the measuring instruments that may be developed with this type of item. The examples are not intended to be representative of all types of multiple-choice questions.

Illustrations of Various Forms and Uses

1. The simplest form of multiple-choice question is that in which a word is given and from a list of other words the student must pick the one that means most nearly the same thing as the given word. The following examples illustrate this form of item:

What is the meaning of the word *coagulate?*

1. dry
2. clot
3. freeze
4. separate
5. become stagnant

What is the meaning of the word *orthodox?*

1. religious
2. heretical
3. accepted
4. aggressive
5. contemporary

In each of the items above, the student must select from the five words listed the one that means most nearly the same as the word italicized in the lead. This type of item is useful for measuring knowledge of vocabulary in any field provided there are a sufficient number of synonyms for the essential words of the subject. In many technical fields vocabulary cannot be measured in this way because there are no synonyms for the technical terms. The same is true of many foreign languages. It is particularly hard to develop a test of French vocabulary in this form since the French language is largely lacking in synonyms. However, where the vocabulary of a field permits this form of test item, it is an effective measuring device.

When vocabulary items of the type above are grouped together, it is customary to tell the student at the beginning of the test that in each question he will first be given a word and then must select from five other words the one which means most nearly the same thing as the given word. Under such conditions the two items given above would be presented in the following form:

coagulate

1. dry
2. clot
3. freeze
4. separate
5. become stagnant

orthodox

1. religious
2. heretical
3. accepted
4. aggressive
5. contemporary

2. Sometimes vocabulary is measured in context. Usually vocabulary is measured in this way in order to determine whether the student understands special uses of words. Such a testing procedure measures the student's understanding of fine shades of meaning rather than his understanding of the general nature of a concept denoted by a word. Examples of items that measure vocabulary in context are given below:

What is the meaning of the word *bulwark* in the following context:

"The school counselor often provides a bulwark between the student and the painful realities of life."?

1. barrier
2. protection
3. parapet
4. support
5. reinforcement

What is the meaning of the word *unassailable* in the following context:

"Smith's reasoning was unassailable."?

1. beyond criticism
2. difficult to attack
3. inconsistent
4. well thought out
5. ineffectual

3. In cases in which it is desired to measure a student's vocabulary but where synonyms do not exist, it is sometimes necessary to give a series of suggested definitions of selected terms. Two examples of this procedure are given below:

siblings

1. children of similar age
2. children born of the same parents
3. blood relatives with a single common ancestor
4. only children
5. children of unmarried parents

conciliation

1. reunification of beliefs
2. settlement of dispute by impartial third party
3. form of bargaining
4. the act of surrendering
5. accepting the terms of an argument

The examples assume that the student has been directed to find among the alternatives the definition of the term given in the lead.

An item in which a series of definitions is given in the alternatives tends to be cumbersome and unnecessarily long and time-consuming. It is sometimes suggested on this account that a reversed form of the above type of item be used in which a definition is given in the lead of the item and in which the alternatives consist as far as possible of single words. The following are examples of items in this form which is sometimes called the reversed-vocabulary form:

The ruler during the youth of a rightful sovereign is called the

1. regent.
2. executive.
3. consort.
4. prime minister.
5. crown prince.

A mathematician who works on insurance problems is called

1. an actuary.
2. a liturgist.
3. a demographer.
4. an underwriter.
5. an agent.

While the reversed-vocabulary form saves the student a considerable amount of reading, there is nevertheless at least one major argument against its use. The argument is simply that there are few cases in life in which a person is confronted with a definition and has to find the word which stands for the same concept. It is much more common for the individual to be faced with a word which he has to explain to another person by defining it.

4. Vocabulary represents only one minor outcome of a whole series in almost any field of teaching. The teacher is usually attempting to develop in the student not only the

ability to name concepts, but also the ability to use those concepts in novel situations. In addition he is concerned with the acquisition of facts which have direct value in the life of the individual.

While memory for facts is a goal of limited importance in most fields, it still has sufficient significance to justify measurement in many situations. The following items illustrate the use of multiple-choice questions for measuring the retention of facts. It should be noted that, in these items, the student either knows or does not know the answer to the problem. There is no way in which he can arrive at the answer by reasoning.

What change occurred in the political status of India in 1947?

1. She was made a mandate of the United Nations.
2. Her people were given their independence.
3. Her people were unified under Moslem rule.
4. She was made a part of the Arab League.

(*Cooperative Contemporary Affairs Test for College Students, Form 1948*, Part I, Item 10. Educational Testing Service, Princeton, New Jersey.)

A device used to change alternating current to direct current is

1. a dry cell.
2. a storage cell.
3. a rectifier.
4. an oscillator.
5. an alternator.

In the earphones of a radio receiver, the diaphragm is caused to vibrate by means of

1. a condenser.
2. an electro-magnet.
3. a transformer.
4. an oscillator.
5. a permanent magnet.

(*Cooperative Pre-Flight Aeronautics Test, Test 5: Radio and Communications and Civil Air Regulations, Provisional Form A,* 1943, Items 12, 39. Educational Testing Service, Princeton, New Jersey.)

The correct sequence of the strokes of a 4-cycle aircraft engine is

1. compression, power, intake, exhaust.
2. intake, compression, power, exhaust.
3. compression, exhaust, intake, power.
4. intake, exhaust, compression, power.
5. intake, power, exhaust, compression.

The material which would most likely be used for piston rings is

1. Babbit metal.
2. bronze
3. nickel steel.
4. aluminum.
5. gray cast iron.

(*Cooperative Pre-Flight Aeronautics Tests, Test 2: Aircraft Engines, Provisional Form A,* 1943, Items 31, 68. Educational Testing Service, Princeton, New Jersey.)

5. Sometimes the retention of knowledge of simple laboratory procedures may be measured as in the following pair of items:

Which of the following is the best way to empty a pipette for use in volumetric analysis?

1. Place the tip of the pipette in the center of the receiving vessel.
2. Place the tip of the pipette against the wall of the receiving vessel.
3. Blow out the last drop.
4. Place the forefinger of one hand on top of the pipette and warm the bulb by clasping it with the palm of the other hand.
5. Blow out the entire contents.

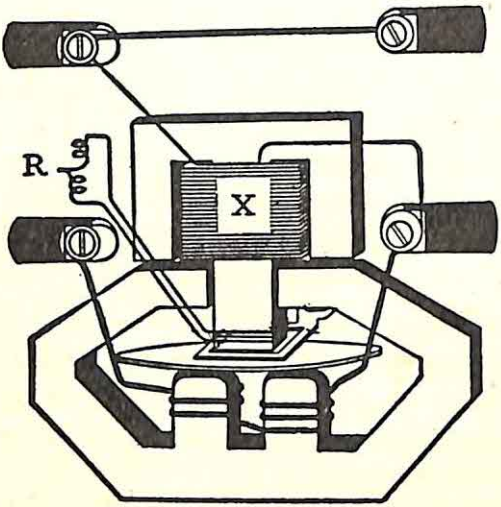
One of the requirements in calibrating volumetric glassware is that

1. the temperature must be that at which the density of water is a maximum.
2. mercury must be used as the calibrating liquid.
3. the temperature must be exactly 20° C.
4. the volume of the piece being calibrated must be compared with the volume of a standard piece of the same kind.
5. the liquid used must leave an unbroken film on the interior when the vessel is drained.

(*A.C.S. Cooperative Chemistry Test in Quantitative Analysis, Form Y*, 1948, Part I, Items 11, 42. Educational Testing Service, Princeton, New Jersey.)

6. Probably the main use of multiple-choice items is for measuring understandings, insights, and appreciations. The following items attempt to measure something more than just the retention of facts and, in most cases, require the student either to see new relationships between facts or to apply principles to relatively novel situations. The following two series of items illustrate the measurement of understanding in technical fields:

Example: [3] Check the statement that indicates the correct answer to each question based on the following drawing.



1. What is the function of the coil marked X?
   a. It produces a magnetic flux proportional to the line voltage.
   b. It exerts a retarding torque on the rotating disc.
   c. It adjusts the instrument for lag.
   d. It exerts a torque proportional to the load.
2. What is the effect on the rotating disc of placing a strong magnet near the rotor?
   a. The speed of rotation is retarded.
   b. The speed of rotation is accelerated.
   c. The speed of rotation is unchanged.
   d. The speed of rotation is either retarded or accelerated, depending on location of the magnet.

[3] In William A. Brownell *et al., The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, pp. 290–291). The University of Chicago Press, 1946.

3. One common way of making the lag adjustment is to adjust the
   a. resistance of R.
   b. position of the drag magnets.
   c. position of the disc.
   d. resistance of the current coils.
4. Which one of the following does NOT change in magnitude when the load is changed?
   a. The field around the potential coil
   b. The currents induced in the rotor
   c. The field around the current coils
   d. The torque exerted on the rotor

Example:[4] To prepare a 100-pound batch of liquid soap meeting the following specifications: (1) two degrees Baumé at 20 degrees centigrade; (2) free alkali not to exceed 0.05 per cent potassium hydroxide; (3) raw materials as follows: cocoanut oil; potassium hydroxide, 47-degree Baumé solution; oleic acid, as needed; perfume, as desired.

1. The main reason for using potassium hydroxide in this process instead of sodium hydroxide is to produce a soap which is
   a. hard.
   b. neutral.
   c. soft.
   d. saponified.
2. One reason why it is necessary to control carefully the alkalinity of the soap produced is that an excess of alkali produces
   a. hydrolysis.
   b. precipitation of grease.
   c. a soap which irritates the skin.
   d. a soap which does not lather.

[4] In William A. Brownell *et al., The Measurement of Understanding*

3. One reason why potassium hydroxide is used instead of calcium hydroxide in making liquid soap is that the latter would produce a soap which is

   a. brown.

   b. acid.

   c. insoluble.

   d. viscous.

4. What is the approximate number of kilograms of potassium hydroxide required to react with 2 kilograms of cocoanut oil if its saponification number is 260?

   a. 52 kg.

   b. 0.26 kg.

   c. 130 kg.

   d. 0.52 kg.

5. Before attempting to make a 47-degree Baumé solution of potassium hydroxide it is necessary to determine the

   a. heat of solution of potassium hydroxide.

   b. degree of ionization of potassium hydroxide.

   c. solubility of potassium hydroxide.

   d. percentage of potassium hydroxide in such a solution.

While the two examples of sets of multiple-choice items given above are derived from technical fields, this does not imply that the only use of the multiple-choice item in measuring understanding is for measuring the ability to apply scientific principles. The examples in the remaining sections of this chapter illustrate the use of multiple-choice items for measuring understanding in a variety of special situations.[5]

(45th Yearbook of the National Society for the Study of Education, Part I, pp. 289–290). The University of Chicago Press, 1946.

7. The next example illustrates the use of multiple-choice items for the measurement of reading comprehension. The student must first read the paragraph and then answer the items, the first of which requires the student to make an inference from the data and the second of which measures understanding of the technique used by the author of the passage:

> If I were founding a university I would first found a smoking room; then when I had a little more money in hand I would found a dormitory; then after that, or more probably with it, a decent reading room and a library. After that, if I still had more money that I couldn't use, I would hire a professor and get some textbooks.

The writer of this paragraph apparently believes that the most valuable part of higher education is provided by

1. the inspiration of great teachers.
2. independent reading.
3. association with other students.
4. close study of a few great books.
5. learning to smoke.

In this passage, the writer achieves a humorous effect by

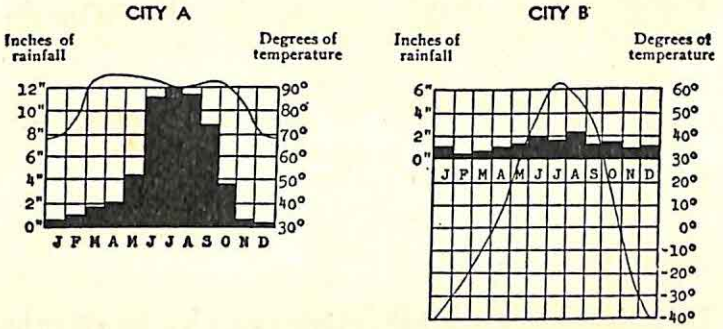
1. making use of understatement.
2. expressing an unconventional point of view.
3. pretending that he has a great deal of money.
4. showing contempt for wealth.
5. implying that books are of no importance.

(*Cooperative English Test, Test C2: Reading Comprehension, Higher Level, Form T*, 1943, Part II, Items 86, 87. Educational Testing Service, Princeton, New Jersey.)

[5] One method that has been suggested for ensuring that a test will

The abilities related to reading which may be measured by such items are varied and include the ability to identify the main ideas in a passage, the ability to interpret characterizations correctly, the ability to understand innuendo, the ability to recognize assumptions made by an author, the ability to recognize the purpose of a speech, the ability to identify the general philosophy of an author, the ability to recognize humor, and many others.

The multiple-choice type of item may also be used to measure skill in the reading of tables, graphs, and diagrams. The following example is drawn from a junior high school social-studies test and illustrates the measurement of skills related to the reading of a graph:



The months are shown by letters— J for January, F for February, M for March, and so forth throughout the year.

Bars show rainfall in inches.

Lines show temperature in Fahrenheit degrees.

---

measure something more than memory for facts is to provide the student with textbooks and other reference materials during the examination. In this way it is hoped that all those taking the examination will have at their disposal the same body of factual information and, in theory at least, can do equally well on those aspects of the test that measure knowledge of facts. Unfortunately, the success of this procedure has not been determined experimentally.

What conclusion can *not* be drawn about the climate of these two cities?

1. Types of clothing needed would be different in city A from city B.
2. The coldest months in city A are also the coldest months in city B.
3. The annual temperature range is greater in city B than in city A.
4. The three months of greatest rainfall in city B are also the three months of greatest rainfall in city A.
5. City B is in the Northern Hemisphere and city A is in the Southern Hemisphere.

About what is the average temperature of city A in February?

1. Between 8 and 10 degrees
2. Between −30 and −40 degrees
3. Between 60 and 70 degrees
4. Between 70 and 80 degrees
5. The graph does not show this.

(*Cooperative Social Studies Test for grades 7, 8, and 9, Form X*, 1947, Part III, Items 9, 10. Educational Testing Service, Princeton, New Jersey.)

8. Multiple-choice questions can be used also to measure the student's understanding of the reasons why certain forms of expression are more effective than others. In the following sample test the student first reads the two versions of the same passage in Column 1 and Column 2 and then answers questions comparing them.

	COLUMN 1		COLUMN 2
A-1	Housing seems to be as much of a problem in the bird world as it is among human beings.	A-2	Housing, a problem in the bird world, is also quite a serious problem among human beings.
B-1	A one-family bird house was inspected by two pairs of bluebirds in a suburban garden the other day.	B-2	The other day two pairs of bluebirds inspected a one-family bird house in a suburban garden.

The inferior version of Section A is poor because

1. emphasis is placed on the wrong part of the idea.
2. the sentence is incomplete.
3. two sentences are punctuated as if they were a single sentence.
4. it is grammatically incorrect.

The better version of Section B is superior because

1. the sentence begins with the subject.
2. parallel structure is used to express parallel ideas.
3. there is a better placing of modifiers.
4. there is more variety of structure.

(*Cooperative English Test, Test B2, Effectiveness of Expression Higher Level, Form T,* 1943, Part I, Items a, b. Educational Testing Service, Princeton, New Jersey.)

9. An interesting variation of the multiple-choice technique has been used for measuring certain outcomes of English instruction. The United States Armed Forces Institute "Test of Effectiveness of Expression," in the series of tests known as *Tests of General Educational Development,* illustrates this kind of technique. In this type of item the subject is presented with connected discourse in which certain words, expressions, and sometimes whole sentences have been underlined. For each underlined section there are several suggested alternatives, one of which may improve the original version. The student is expected to select the version which seems best to him. This type of item is commonly found in tests which are supposed to measure "effectiveness of expression." As a matter of fact, such tests do not measure effectiveness of expression at all, but they do measure the student's ability to recognize the most acceptable way of stating certain ideas and concepts. Note also in the following example that reading comprehension is one of the most important factors in solving the problems:

While it is likely that our competitors have produced a small quantity of a new plastic similar to our own, it would be dangerous at this time to assume that they would be four years behind
1
us. It would be much more reasonable assuming that they have
2
geared their plants to produce at a rate similar to our own. . . .

1. a. are
   b. were
   c. will be
   d. correct as it stands

2. a. assumption
   b. to assume
   c. to be assumed
   d. correct as it stands

10. While the multiple-choice type of item has been found to be an effective device for measuring various reading skills, there has been a tendency in recent years to measure understanding in specific subject-matter fields through tests of reading comprehension in those fields. The *Tests of General Educational Development* of the United States Armed Forces Institute were attempts to measure understanding in broad areas of study, and the Cooperative Test Service published a similar series of achievement tests. The following illustration shows how the interpretation of reading matter may be used to measure understanding in the social studies:

> Private enterprise motivated by the hope of profits does not increase production until effective demand is actually in sight. Since a large part of the potential effective demand must consist of wages and salaries, the demand does not increase substantially until *after* production has been increased. Privately owned companies cannot be expected voluntarily to increase the volume of wages paid to their workers merely to increase the demand for their products, because without corresponding action by all other producers they could not expect their receipts to go up as much or as rapidly as expenses.

The writer implies that private enterprise as a whole would be benefited by

1. less government interference.
2. a high volume of wage payments.
3. a reduction in the expenses of producers.
4. the competition of government-owned enterprises.
5. a high rate of profit.

What reason would the writer be likely to give for the rapid increases in war production during the war?

1. Increased supplies of raw materials obtained through reverse lend-lease
2. The employers' desire to defend the free-enterprise system
3. Control of raw materials by the War Production Board
4. The patriotism of the employees
5. Government purchasing

(*Cooperative General Achievement Tests, Revised Series, 1, of Form X*, 1947, Part II, Items 12, 15. Educational Testing Service, Princeton, New Jersey.)

The value of reading-comprehension items for measuring achievement has not been well established, and there is some question as to whether understanding cannot be measured more economically by simpler techniques. The chief argument for their use is that by providing a reading passage it is possible to require the student to interpret the material without requiring him to recall specific details of the subject matter involved. It may also be noted that although it has been shown that reading-interpretation tests in specific subject-matter fields may have value for predicting subsequent grades in those fields, this does not necessarily mean that they are valid for measuring achievement.

11. A variation of the multiple-choice question which is commonly used for measuring rather superficial understanding is the classification item. The clearest way to describe this type of item is by presenting an example:

After Alice had studied meal planning, she asked her mother if she might plan and prepare dinners at home. Her mother said she could, but pointed out that careful planning would be necessary because Alice got home only an hour before the time that the family usually ate dinner.

Below are listed four menus that Alice planned, followed by a series of questions about the menus.[6]

MENU 1

Rib Roast of Beef
Scalloped Potatoes
Buttered Cabbage
Bread Butter
Fruit Cup
Coffee or Milk

MENU 2

Baked Macaroni and Cheese
Boiled Potatoes
Banana and Nut Salad
Biscuits Butter
Apple Jelly
Cottage Pudding
Coffee or Milk

MENU 3

Creamed Chipped Beef on Toast
Frozen Peas Cauliflower
Mixed Green Salad
Bread Butter
Chocolate Nut Pudding
Coffee or Milk

MENU 4

Clear Tomato Soup
Cabbage Salad
Carrot Strips Celery
Bread Butter
Baked Apples
Coffee or Milk

Which meal provides the best variety of food values, colors, flavors, and textures?

Which meal contains the most foods of similar textures?

(*Cooperative Test in Foods and Nutrition, Form Y*, 1948, Part II, Items 119, 122. Educational Testing Service, Princeton, New Jersey.)

[6] Only two of the questions are given for illustrative purposes.

In items of the type shown above, the student must answer each question by selecting the category (in this case menu) that best fits a given classification. It would be possible in the example above to add a whole series of questions based upon the same data. Problem situations of this type may have twenty or more items based on the same data.

One major limitation of the classification item which, in the view of the present writer, is sufficient to prohibit its use in most situations is that the four choices do not function in most of the questions. For example, if the question were asked in the above problem, "Which menu probably has the lowest calorific content?" alternatives 2 and 3 would not enter into the choice of a solution. Usually, classification items resolve themselves into 2-choice items and are therefore rather uneconomical in view of the quantity of material that must be read before they can be solved. Another limitation of the classification type of item is that it measures knowledge at a superficial level. Often it measures little more than the kind of knowledge that is acquired by rote learning. Only rarely does it measure understanding. The outcomes it most often measures are trivial.

12. The matching item is a special form of the multiple-choice item and one which is generally considered to have relatively little value as a measuring device. In order to appreciate the limitations of this type of item, the following example is presented as a basis for discussion:

> *Directions:* The following problem presents a column listing the names of several important scientists and a column listing some important theories and scientific discoveries. Find the discovery or theory that is associated with each name in the left-hand column and write the letter of that theory or discovery opposite the name:

1. Linnaeus _____
2. Harvey _____
3. Macleod _____
4. Mendel _____
5. Koch _____
6. Pasteur _____
7. Pavlov _____

a. Treatment for syphilis
b. The discovery of the nervous impulse
c. The discovery of the circulation of the blood
d. Theory of the inheritance of acquired characteristics
e. The motivation theory
f. Crucial experiments on the sources of living organisms
g. Conditions under which the conditioned reflex occurs
h. Initial discovery of cell structure
i. The mechanism of the transmission of malaria
j. The theory of natural selection
k. The discovery of male sex hormone
l. The mechanism of heredity

If the reader will attempt to solve this problem he will begin to see its deficiencies. If Linnaeus is recognized as a figure out of the early history of science, the student will immediately know that many of the relatively recent discoveries could *not* be associated with him. In other words, only some of the items in the right-hand list offer possible

alternatives to the student who knows little about the field. The other alternatives represent just so much dead wood and time spent in unnecessary reading. The dead wood can be eliminated from each item only by converting it into a multiple-choice form, that is to say, for example, by asking the question, "Which one of the following discoveries was made by Linnaeus?" and then giving a series of alternatives each of which may plausibly be attributed to Linnaeus. The matching item given above would be a much better measuring instrument if it were converted into seven multiple-choice test questions.

Now consider the following problem and note the cues that can be used in solving it:

Directions: The following problem presents a column listing the names of several important persons in American politics and a column listing several political doctrines, acts, and organizations. Choose the item in the right-hand column which is most closely associated with each name and write the number of that item opposite the name.

a. George Marshall ____
b. Robert Wagner ____
c. Cordell Hull ____
d. Herbert Hoover ____

1. Reconstruction Finance Corporation
2. Appeasement
3. Reciprocal trade agreements
4. Labor Relations Act
5. European Recovery Program
6. Government-supported housing program
7. Parity prices for farmers

It will be noted that this problem, like most matching problems, measures a very superficial level of understanding. In a few exceptional circumstances it is possible to make matching problems that measure something more than memory for facts, but something can usually be gained by converting them into items of the multiple-choice type.

A special case of the matching item is found in tests in which the student is requested to arrange a series of events in chronological order. Tests of this kind present difficulties of scoring which cannot be overcome by any simple device. What the teacher wants to know is the extent to which the order given by the student corresponds to the correct order. While this is a simple problem of rank-order correlation it is not one which a teacher can easily handle. There are also other factors which should discourage the teacher from using this form of item and of particular note is the relative unimportance of the outcome which it measures.

13. The examples of multiple-choice items for measuring understanding given up to this point have been presented as groups of items based on a common setting. However, it should be noted that single items may also measure understanding. The following items are presented as illustrations:

A thick-glass beaker breaks more easily than a thin one when hot water is put into it because

1. thick glass holds the heat better.
2. thick glass is more brittle.
3. thick glass expands more rapidly.
4. the inside of the thick glass expands before the outside is warmed.
5. glass is a good heat conductor.

Occasionally an entirely white corn plant occurs in a field of corn. This plant never reaches maturity because

1. it cannot manufacture food.
2. hot weather kills it.
3. insects will attack it.
4. it cannot take water from the soil.
5. it has no root hairs.

(*Cooperative General Culture Test, Revised Series, Form U*, 1944, Part IV, Items 5, 7. Educational Testing Service, Princeton, New Jersey.)

Which of the following best explains why ice is an excellent refrigerant?

1. It absorbs much heat while melting.
2. It can be placed in a refrigerator at sub-freezing temperature.
3. It can change directly from a solid to a gas.
4. It cools by evaporation.
5. It is an excellent conductor of heat.

Why should the grounded ends of lightning rods extend down into moist subsoil?

1. Heat from lightning is cooled by water.
2. Water draws lightning flashes into the soil.
3. Moist soil is a good conductor.
4. Moist soil is a poor conductor.
5. Dry soil is a poor insulator.

(*Cooperative General Science Test, Revised Series, Form X*, 1947, Part II, Items 25, 29. Educational Testing Service, Princeton, New Jersey.)

During the decade just before the Civil War, commerce shifted away from the port of New Orleans and toward the port of New York. Which of the following best accounts for this shift?

1. Perfection of the river steamboat
2. East-west railroad construction
3. Anti-slavery feeling among Northern farmers
4. Decline of cotton as an important export
5. Export taxes levied by the state of Louisiana

Americans have shown the greatest amount of architectural originality in the construction of

1. churches.
2. large stadiums.
3. large business buildings.
4. state-capitol buildings.
5. museums and libraries.

(*Cooperative American History Test, Revised Series, Form Y*, 1948, Part II, Items 8, 19. Educational Testing Service, Princeton, New Jersey.)

Italy's object in making alliances before both World Wars was

1. to advance the cause of democracy in Europe.
2. to gain security against French invasion.
3. to re-establish the balance of power in Europe.
4. to acquire the necessary strength to resist a British naval blockade.
5. to further her territorial interests in the Mediterranean region.

The teachings of Confucius have influenced Chinese history by

1. persuading the Chinese to worship only one God.
2. popularizing the idea that the end justifies the means.
3. slowing the rate of adjustment to modern conditions.
4. encouraging the movement for national unity.
5. sponsoring vocational schools for the peasants.

(*Cooperative Modern European History Test, Revised Series, Form X*, 1947, Part I, Items 11, 40. Educational Testing Service, Princeton, New Jersey.)

What is the most accurate method for determining the *p*H of the urine?

1. Colorimetric comparison with a solution of known *p*H
2. Titration with a standard solution of an acid or of a base
3. Measurement of the E.M.F., using a glass electrode
4. Titration with a standard solution of an oxidizing or of a reducing agent
5. Observation of the color change of a "Universal" indicator

Representative 24-hour urine samples of a human subject on a normal diet were analyzed. The protein content of the diet was then considerably increased and urine analyses were again made. Which of the following results might be expected?

1. A significant rise in the creatinine nitrogen
2. A considerable rise in the *p*H of the urine
3. A considerable increase in ammonia excretion
4. A significant rise in the excretion of uric acid
5. Slight changes in titratable acidity

(*Cooperative Biochemistry Test, Form X*, 1947, Part II, Items 125, 128. Educational Testing Service, Princeton, New Jersey.)

What was the immediate value of the purchase of Louisiana?

1. The United States made its claim to Florida secure.
2. Western settlers were guaranteed the right to navigate the Mississippi River.
3. The region became the center of the cane and beet sugar industry.
4. The region abounded in fertile farms and rich pastures.
5. The gold and silver mined in the mountain states was worth many times the amount paid for the whole territory.

One of the most important reasons for the growth of European imperialism in the second half of the nineteenth century was the

1. need for more lands to relieve the crowding in European countries.
2. desire to spread the Christian religion.
3. demand for new markets for manufactured products.
4. need for additional land for grants to the nobility.
5. fear of American sea power.

(*Cooperative General Culture Test, Revised Series, Form X*, 1947, Part II, Items 20, 50. Educational Testing Service, Princeton, New Jersey.)

The reader should study each of the test items above and note that while each *may* measure understanding, there is a some possibility that it will not. If the specific problem given in an item has been previously discussed in class, and if the solution has been given, the item will test the student's ability to remember the solution. In such a case, the item will measure the retention of a fact and not basic understanding

of the field. It is possible to determine whether an item measures understanding only if the educational background of the student is known.

SPECIAL TYPES OF RESPONSES

In recent years certain responses such as "none of the above" and "all the above" have appeared frequently in objective examinations. Since these alternatives should be used only with the greatest caution, some consideration must be given to their merits and limitations. The answer, "none of the above," has been used most frequently in tests of mathematics. The object of using such an alternative is to make the examination as similar as possible to a free-answer type of instrument while retaining the multiple-choice form. When it is used appropriately, the answer, "none of the above," is the correct answer in a proper fraction of the items. There is as yet little evidence to show that the use of this alternative produces an examination which functions in a similar way to the free-answer examination.

In general, it is believed on the basis of rather inadequate evidence that the main use of the alternative "none of the above" is in tests of outcomes in mathematical areas. However, there is one exception to that statement in the case in which the solution to a problem is obvious. In such instances it is possible to determine by means of an objective-test item whether the student can recognize certain wrong solutions as being wrong, but it is not possible to determine by this method whether the student knows the right solution. In such cases it is sometimes justified to use "none of the above" as a right answer, provided that it is used also as a wrong answer in certain other items.

The answer, "all the above," can be justified only rarely.

What happens too often when it is used is that most pupils read the first of the alternative solutions, and noting that that solution is correct do not go on to read the other alternatives. Consequently, in such items, there is a marked tendency for the first answer rather than the fourth answer ("all of the above") to be marked as correct. A rather unsatisfactory way of overcoming this difficulty is to change "all the above" to "all of the following" and to put the latter response as the first suggested response. There are, however, other major arguments against the use of this alternative. When it is used, it means that the test writer must abandon one of the great merits of the multiple-choice problem, namely, the fact that it can require the student to discriminate among various possible solutions to a problem and eventually to arrive at a decision concerning which one is best. In general, it is suggested that the responses "all the above" and "all of the following" be avoided.

Chapter Five

Rules for Constructing Multiple-Choice Test Questions

DURING THE LAST FIFTEEN YEARS persons engaged in the production of objective tests have developed a series of rules that have become widely accepted as a formula for building good best-answer items, once a test plan has been properly prepared. It must be recognized that these rules are not based on careful experiment, but have been evolved partly on the basis of the experience of test technicians with the way in which test items work out in trial runs and partly on the basis of an introspective and psychological analysis of what is done in the solution of a test problem.

There is a real need for the systematic study of the value of these rules, and until the outcome of such a study is available they must be presented as a formula based only on the judgments of experts.

The rules will be stated in three groups. First, there will be a set of general rules. Second, there will be a set of rules for developing the lead or stem of the item; and third, there will be a set of rules for developing the suggested solutions. Many of the rules may seem to be so obvious that the reader may wonder why they are stated, but a review of examina-

tions given in schools and colleges reveals that even some of the most obvious rules are frequently broken. At least one reason for this fact is that many examinations are given mainly for the purpose of giving a grade and not for the purpose of evaluating student development. Also, it should be noted that the rules apply to an ideal test item, and it is hardly possible to develop such an item.

While a few of the rules given in this chapter are substantially the same as those given for constructing true-false and completion items, they are restated here together with examples of best-answer items.

GENERAL RULES

1. The item as a whole should present a problem of a kind that will enable the student, if he can solve it, to show evidence of the attainment of an important goal.

Many items in many tests present trivial problems and do not give the student opportunity to provide significant evidence of whether or not important goals have been achieved. Consider, for example, the following question:

Who was Vice-President of the United States during the first administration of Franklin Delano Roosevelt?

1. John Garner
2. Cordell Hull
3. Henry Wallace
4. Alben Barkley
5. Stephen Early

This question asks for a relatively trivial fact concerning the Vice-Presidency, and certainly, if there were to be only one question asked in an examination on this topic, a more

significant one could be found. Contrast the item above with the following one, which attempts to determine whether the student has acquired an important piece of information about the Vice-Presidency.

> Someone has jokingly referred to a prominent Federal official as "His Superfluous Excellency." Who most nearly fits this description in normal times?
>
> 1. Vice-President
> 2. Speaker of the House of Representatives
> 3. Director of the Budget
> 4. Treasurer of the United States
> 5. Secretary of the Interior
>
> (*Cooperative American Government Test, Revised, Form X*, 1947, Part I, Item 20. Educational Testing Service, Princeton, New Jersey.)

In appraising a test which a teacher has developed, every problem must be reviewed, and the question must be asked: "Does the student's response to the problem provide evidence of the attainment of an important educational goal?"

Related to this matter is the fact that, in many tests, the most difficult items are difficult, not because they cover applications of principles which require careful thought and analysis, but because they cover minor elements which are of significance only to the expert. If it is necessary to develop difficult items, difficulty should in most cases be achieved, not by making the item trivial, but by introducing an element of novelty.

It should be noted that it is not the content in itself, but rather how the content is handled that determines whether or not an item is related to a significant domain of behavior. Examine the following pairs of items which cover similar content areas and contrast in each pair the triviality of the

first item with the importance of the concept covered by the second item:

TRIVIAL ITEM ABOUT AN IMPORTANT AREA OF CONTENT

Which Amendment to the Constitution is concerned with the right to bear arms?

1. 1st
2. 2nd
3. 3rd
4. 4th

SIGNIFICANT ITEM COVERING THE SAME CONTENT AS THE ITEM AT THE LEFT

The Bill of Rights guarantees to the people the right to keep and bear arms. One of the basic concepts underlying this amendment is that

1. armed strength should be in the hands of the people.
2. hunting is an important national activity.
3. the free state should place few restrictions on the individual.
4. arms are symbols of freedom.

ITEM MEASURING MEMORY FOR DETAIL

Which of the following statements is derived from President Franklin D. Roosevelt's first inaugural address?

1. "The only thing we have to fear is fear itself."

ITEM MEASURING UNDERSTANDING

In his first inaugural address President Franklin D. Roosevelt referred to the need of casting out the moneychangers from the temple. He was referring to the need for

2. "Only when the night is darkest can we see the stars."
3. "Each must receive his just share of the fruits of the Earth."
4. "The people must regain the power that rightfully belongs to them."

1. curbing the power of Wall Street.
2. monetary reform.
3. revising the gold standard.
4. restricting credit buying.

2. Items which attempt to measure insights must include an element of novelty.

It is fairly evident that if understanding rather than rote memory is to be measured, it is necessary to present the student with problems other than those he has already solved. By giving the student a series of familiar problems, the teacher may be measuring nothing more than the ability of the student to remember solutions that he has previously seen, and such items do not necessarily measure thinking skill at all. It is psychologically sound to say that thinking occurs in those situations for which the individual has no ready-made response. Consequently, in order to measure thinking skills, it is necessary to present the student with problem situations for which he has no immediately available response, that is to say, problems in which there is a combination of elements which have not previously been encountered together.

It should be noted that the teacher is in the best position for judging whether a problem situation has novelty for the student. A problem which may be highly novel to one child is familiar to another. Novelty depends largely on the experiences to which a student has been exposed. It has been pointed out in an earlier chapter that one of the main

reasons for outlining the content of a course, as a foundation for the development of tests, is that such outlining permits the test technician to exercise some control over the element of novelty in the test items. A note of warning should be introduced at this point. Too much novelty in a test problem usually results in a so-called "trick question." Trick questions represent merely a misuse of the element of novelty and are never justified in any rational plan of evaluation.

3. The language used in stating the problem should be appropriate to the subject matter.

This rule is particularly important in technical fields where special idioms are used, but it should be kept in mind in most fields. Problems to measure outcomes of sixth-grade social studies should be stated in the kind of language that all sixth graders will understand. Too often examination problems are phrased in highly academic language which is inappropriate to the given subject matter from the point of view both of the idiom and of the level of difficulty of the vocabulary and sentence structure. Particular care must be taken in measuring the outcomes of shop courses, for it happens all too frequently that questions covering shop practices are expressed in academic language of a type which shop workers do not use. The only sound criterion of the appropriateness of the language is whether the individuals who take the test understand the problems, even though they may be unable to solve them. In test writing, as in other fields of authorship, quality of writing must be judged largely in terms of whether the written material conveys precisely and unambiguously the ideas it is supposed to convey.

Considerable help can be obtained in judging the difficulty of vocabulary by using the Thorndike-Lorge semantic count.[1] This work enables the writer to determine the fre-

[1] Thorndike, Edward L., and Lorge, Irving. *The Teacher's Word Book of*

quency with which a large number of words are encountered in common reading material. However, it should be noted that vocabulary difficulty is only one of the factors which determine the level of difficulty of a passage. Some examples of leads appropriately and inappropriately worded are given below:

Inappropriate Leads	More Appropriate Leads
What is the function of the clapper box on the shaper?	The clapper box on the shaper is used for . . .
The main factor which dedetermines the speed of operation of the acetylene welder is the . . .	The speed of work of the acetylene welder depends mainly on the . . .
What operation must be undertaken first in sharpening a hand saw?	The first thing to do in sharpening a hand saw is to . . .

Sometimes items are prepared in an unsuitable form as a result of a tendency to present in words items involving space concepts. The following item from a test of beginning drawing is typical of this error:

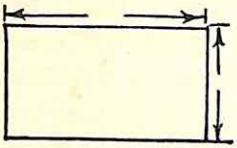
Where should dimension lines be placed on a front view?

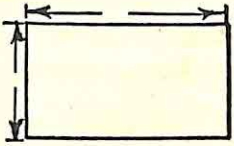
1. At the top and at the right
2. At the top and at the left
3. At the bottom and at the right
4. At the bottom and at the left

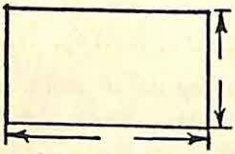
This problem may cause some students unnecessary difficulty because it requires the ability to visualize in addition to knowledge of a fact. Contrast the statement of the problem given above with the following version:

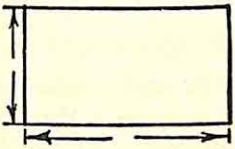
30,000 Words. Bureau of Publications, Teachers College, Columbia University, New York, 1944.

In which one of the following drawings are the dimension lines correctly located in a front view?

1. 

2. 

3. 

4. 

4. The items in the test should be independent of one another, and the information supplied in one item should not give away the answer to another.

This rule is not likely to affect more than a few items in most tests. It is the least important general rule of those given. However, before a test is reproduced, but after all the items have been written, the test as a whole should be reviewed to insure that the information given in the leads of some of the items does not help the student to solve certain other items. Examples of items in which the answer to one is found in the lead of the other are given below:

What method did Banting and his associates use in the original extraction of insulin from the pancreas?

1. Solution in alcohol
2. Crystallization from aqueous solution
3. Precipitation
4. Rectification
5. Differential diffusion

In which gland is insulin formed?

1. Pancreas
2. Liver
3. Spleen
4. Adrenal
5. Pituitary

If the items above were given in a test, the student could answer the second on the basis of the information given in the first. If it were desired to include both items, the first could be revised by cutting out all unessential information. The lead would then read, "What method was used in the original extraction of insulin?"

RULES FOR STATING A PROBLEM

It is possible for an item to be built around all the rules stated above and yet for it to fail to be an efficient measuring device. Certain rules should be observed in the statement of the problem if the item is to meet acceptable standards. These rules are stated below.

1. The lead of the item must present a single central problem.

Too frequently the lead of an item presents no problem at all; or it presents too many problems. For example, the following items are typical of those in which there is no central problem:

Pumping oil to the surface of the earth

a. increases the angular velocity of the earth.

b. increases the kinetic energy of rotation.

c. increases the angular momentum of the earth.

d. decreases the angular velocity of the earth.

Geographic factors and natural resources

1. determine the nature of the culture in any area.

2. are usually unrelated to the nature of the culture in any area.

3. affect the material aspects of the culture but not its social aspects.

4. exert a limiting and permissive but not a determining effect on culture.

In both the problems above the teacher probably was trying to find out whether the student had some particular understanding, but failed, in each instance, to present a definite problem. These problems could be restated in the following form:

To what extent, if at all, does the pumping of oil to the surface at the equator affect the angular velocity of the earth?

To what extent, if at all, do geographic factors and natural resources determine the nature of the culture in an area?

When the problems are thus restated they become definite, and by the time the student has read the leads he

knows just what the problems are that he is supposed to solve. He can then go on to consider the merits of the various solutions for them. The reader should note that the second problem is probably too broad in scope and should therefore be broken down into a series of more limited problems.

One method of determining whether there is a central problem in the lead of an item is to cover up the alternatives, and thus to see whether the problem, standing alone, is still intelligible. Every multiple-choice problem ought to pass this test. In other words, every multiple-choice problem should be usable as a free-response item.

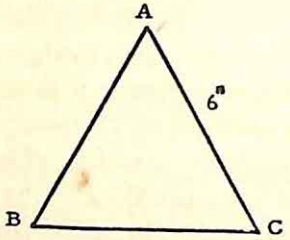
A common form of multiple-choice item which completely lacks a central significant problem starts with the words, "Which one of the following statements is true?" This kind of lead and any variations of it should be avoided. It is evident that a problem which merely asks the student to determine which statement is true is nothing more than a series of true-false items and therefore loses the advantages which the multiple-choice form possesses.

2. ***The problem must be accurately stated.***

In many test questions the problem is stated in such vague terms that the student must try to determine what the examiner has in mind before he can attempt to solve it. It is fairly obvious that the student should not have to guess what the teacher is thinking; yet it is a temptation for the teacher, in writing examinations, to say to himself, "They (the students) will know what I mean even if it is not absolutely clear." Often, however, the student makes a very poor guess at what the teacher has in mind.

The essential information so often omitted from test items is a statement of some assumption that must be made in order to solve the problem. The following items illustrate this common inadequacy:

INACCURATELY STATED PROBLEM



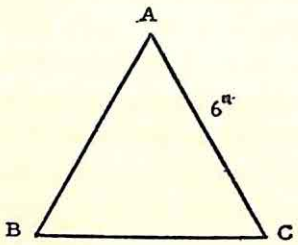
What is the area of this equilateral triangle?

1. $9\sqrt{2}$ sq. in.
2. $3\sqrt{3}$ sq. in.
3. $9\sqrt{3}$ sq. in.
4. 18 sq. in.
5. 27 sq. in.

The upward forces on an airplane must be approximately equal to the

1. total weight of the plane.
2. weight of the air displaced.
3. weight of the useful load of the plane.
4. momentum of the plane.

ACCURATELY STATED PROBLEM



What is the area of this triangle?

1. $9\sqrt{2}$ sq. in.
2. $3\sqrt{3}$ sq. in.
3. $9\sqrt{3}$ sq. in.
4. 18 sq. in.
5. 27 sq. in.

When an airplane is flying in a horizontal direction, the upward forces on the plane must be approximately equal to the

1. total weight of the plane.
2. weight of the air displaced.
3. weight of the useful load of the plane.
4. momentum of the plane.

Inaccuracy in the statement of the problem is often responsible for much of the unconscious humor that appears

in objective examinations. While it is quite clear what the author of the following item had in mind, there is a discrepancy between what he intended to say and what he did say.

Which type of insurance is best for a man wishing to save money to put his baby son through college?

a. Term
b. Limited-payment life
c. Whole life
d. Endowment

Sometimes items fail to function in the way expected because they are ambiguous. For example, the following item appeared in a test in anthropology:

In what percentage of the peoples of the world has polyandry probably existed?

1. 1
2. 5
3. 10
4. 20

It is not clear whether the author of the above item was concerned with living peoples of the world or with all peoples of the world that have ever existed. It should be noted that ambiguities in items can be found best by having them reviewed by several people before they are used.

3. The problem should not measure the ability to understand complex sentence-structure except when it is desired to measure that particular ability.

The following item illustrates a gross violation of this rule:

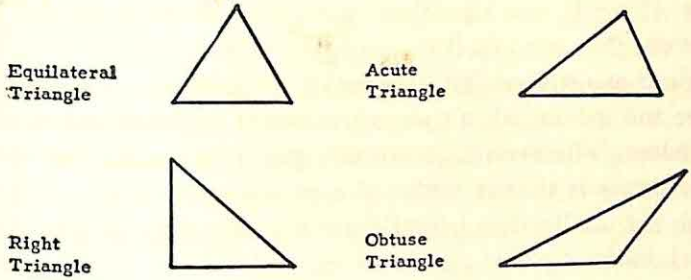
> Each alternative contains *two* characteristics *either* of which may be true of a culture. Which one includes two characteristics which are almost certain to occur together in the *same* culture?
>
> 1. Belief in supernatural causation; little control over natural environment
> 2. Commercial development; little cultural lag
> 3. Agricultural development; rapid institutional change
> 4. Paternalistic government; extensive civil liberties
> 5. High rate of invention and discovery; little cultural lag

In this item the student may have difficulty in finding out what the problem is in the first place. The following revision shows how the problem may be stated in much simpler terms:

> Which one of the following characteristics of a culture is almost certain to accompany a belief in supernatural causation?
>
> 1. Little control over natural environment
> 2. Commercial development
> 3. Extensive civil liberties
> 4. Economic security
> 5. Religious persecution

The following example from a test in high-school mathematics also illustrates unnecessary complexity in the presentation of an item:

Given:



These triangles are all named according to the angles they contain. Any triangle would fall into one of these types. What can you say about the necessity of one or more acute or obtuse angles being present in any plane triangle?

a. All plane triangles contain two acute angles.

b. All plane triangles contain one obtuse angle.

c. All plane triangles contain two obtuse angles.

d. All plane triangles contain three acute angles.

The item above does little more than attempt to find out whether the student knows that in a plane triangle there must be at least two acute angles. The student's knowledge of this fact can be measured much more economically by the following item:

How many acute angles, if any, are *necessarily* found in any triangle?

1. 2 or 3

2. 2

3. 3

4. 0

4. In most cases a problem should contain only material that is relevant to its solution.

This rule is really a specific case of the previous rule. However, there is one situation in which this rule should be broken; that is when it is desired to determine whether the student can discriminate between information that is relevant and information that is irrelevant to the solution of a problem. The reason commonly given for including irrelevant data is that it makes the problem more interesting. This rationalization is really a confession that the problem is trivial in the first place. An example of a problem stated with unnecessary data and the same problem stated in its simplest form is given below:

> Johnny goes to the store to buy some potatoes. If the price of potatoes is 6¢ per pound, what does he pay for 12 pounds?
>
> What is the price of 12 pounds of potatoes at 6¢ per pound?

In the first version no advantage is gained by introducing Johnny into the problem. Indeed, so far as the problem is supposed to measure arithmetical skills, the second version is better because it places less emphasis on reading skill. It is doubtful whether the problem is made more interesting by the introduction of a person.

The following item illustrates the same point:

> During the summer of 1946 much bitter feeling was aroused over the issue of removing price controls. The main argument of those who favored the lifting of price controls was that it would
>
> 1. stimulate production.
> 2. increase the national income.
> 3. increase the income of the farmer.
> 4. lower prices.

It is evident that the first sentence in this item is quite unnecessary and can be deleted without loss. The "moral" is that in a well-planned educational program the problems the pupils are learning to solve are necessarily interesting to them, and do not need to be garnished.

5. The problem should be stated in a positive form.

There are two main reasons why the lead of an item should, whenever possible, be stated positively rather than negatively.

First, it is evident from item analyses that many students respond to problems stated in a negative form as if they were stated in a positive form. The reason for this response is not clear; but it is known that items that have negative statements in the leads tend to have rather low validity. In order to overcome this fault, it is usual to underline or to capitalize such words as *not* and *never,* and such prefixes as *un* in *un*desirable, in order to bring them to the attention of the reader. This practice, however, does not fully eliminate the tendency for such negative words and syllables to be overlooked.

Second, leads containing negative statements are likely to be trivial in character. In reading an item like the following, the teacher will naturally feel that the problem would have greater significance if it were phrased in positive form:

To a sociologist world moral integration does *not* depend on which of the following?

1. Increased international communication
2. International control of atomic energy
3. Attachment to world-culture symbols
4. Appreciation by the people of the world of the facts of human nature

The reader will almost certainly have the impression that this problem would be more significant if it were stated in positive form. Surely it is more important that a student know the positive factors making for moral integration than that he know some of the factors that bear no relationship to it.

6. If an item requires the student to express an opinion or value judgment, it should in most cases ask the student to express, not his own opinion, but that of an authority specified in the item.

In most objective examinations the teacher is not specifically concerned with the inner feelings of the student. In general, the best-answer type of objective test has not been widely used for studying the individual student's own value judgments for, as the name implies, each item in it should have a right answer or an approved or acceptable answer. However, the best-answer type of question has been widely used for measuring the student's knowledge and understanding of the value judgments made by others. As was pointed out previously, one cannot ask the question, "What is the best method of providing economic security for miners in their old age?" since the answer cannot be scored. However, at the time of writing it is possible to score the answer to the question "What does John L. Lewis believe to be the best method of providing economic security for miners in their old age?"

There are a few situations in which it is permissible to use multiple-choice questions for the appraisal of the student's own value judgments. Such situations are those in which the student is unable to choose deliberately the answers which the teacher would prefer. Probably the best-known instance of this type of evaluation instrument was developed by Abbott and Trabue, who selected a large number of poems that are recognized as works of considerable

merit. These poems were modified in ways that destroyed their original worth in varying degrees. Some were converted into sentimental versions of the original, some into prosaic versions. Three new versions of each poem were prepared and, together with the original version, were submitted to a number of experts including poets, literary editors, critics, and professors of literature. If these experts did not agree that the original was the best of the four versions, the series was discarded. In some cases, it seems, a change in a poem actually improved it.

After four versions of more than one hundred poems had been so examined, two short series of thirteen poems each were selected for the two forms of the final test. Each of these series represented material of graded difficulty from Mother Goose up to Milton or Browning. In the test itself each page was devoted to the four versions of a single poem and the subject was instructed to "read the poems, A, B, C, D, trying to think how they would sound if read aloud. Write 'Best' on the dotted line above the one you like best as poetry. Write 'Worst' above the one you like least."

These tests were given to a large number of children in elementary and high school and to students in college, in order to determine how accurately young people of different ages were able to evaluate the relative merits of the four versions of a given poem. It is obvious that the student whose taste corresponded most closely with that of the experts would be able to choose correctly the best versions of the thirteen poems in each series and would thus score thirteen points.

RULES FOR DEVELOPING SUGGESTED SOLUTIONS

1. The right solution should be unquestionably right, and at least two persons should review the item and agree upon the correct solution.

This rule may seem to be so obvious that the reader may wonder why it is stated, but the fact is that one of the major defects of most objective examinations is that they include problems for which no completely correct solution is given. Sometimes the lack of a correct answer is due to the fact that the test writer failed to indicate the units in terms of which the answer is to be stated. For example, a teacher wrote the following problem:

If a man walks six miles in two hours, how fast is he walking?

1. 2
2. 3
3. 6
4. None of the above

The teacher who wrote this item intended the correct answer to be "3 miles per hour," but since the unit "miles per hour" is not included in the alternatives the student must decide whether the teacher really meant "3 miles per hour" in the second alternative or, if not, whether the last alternative is the correct one.

A teacher often accepts a partially correct answer because it is the best of the given alternatives. This practice should be avoided, since it is disconcerting to the student and likely to arouse considerable and unnecessary argument. Fre-

quently a partially correct solution can be made wholly correct by modifying the lead, as in the following example: (In each of the two items the first alternative was keyed as correct.)

What was the purpose of the Taft-Hartley law?	One purpose of the Taft-Hartley law was to
1. To eliminate communists from positions of leadership in unions	1. eliminate communists from positions of leadership in unions.
2. To limit the size of unions	2. limit the size of unions.
3. To reduce the initiation fees and other dues of union members	3. reduce the initiation fees and other dues of union members.
4. To curb inflation resulting from certain union practices	4. curb inflation resulting from certain union practices.

2. The suggested wrong answers should represent errors commonly made by the students who are to be tested, not popular misconceptions held by people in general.

The reasons for this rule are fairly obvious. In most courses the students at the beginning hold certain popular misconceptions related to the subject-matter field. However, most of these popular misconceptions will be dispelled before the course is completed. This does not mean that at the end of a course a student does not have any misconceptions about the subject. He does; but his misconceptions are much more likely to be errors of technical judgment than errors resulting from almost complete ignorance. From this rule it follows that the person who is best equipped to write

the suggested answers to a problem, and for that matter to write the problem itself, is the teacher who knows the kinds of difficulties students encounter and the ways in which they misinterpret the experiences the course provides. Many would say that the teacher with recent classroom experience is the only person who is adequately equipped to develop evaluation procedures. There are, of course, other reasons, such as those discussed in the preceding pages, why the teacher should play the central role in all evaluation procedures.

3. The suggested answers should be as brief as possible. In order that a minimum of reading skill may be required for solving the problems, it is desirable to make the suggested answers brief. The following item illustrates unnecessarily long alternatives:

One group which opposed the adoption of the United States Constitution was the

1. Southern slaveowners, who feared immediate termination of the slave trade.
2. debtor element, who opposed restriction of state powers over the currency system.
3. commercial and trading interests along the Atlantic Seaboard, because they wished to continue state regulation of interstate trade.
4. well-to-do farmers east of the Mississippi, who feared the power of a distant central government.

This item includes a great deal of unnecessary material in the alternatives. It would be materially improved if it were rewritten in the following way, with the omission of much of the unnecessary data:

One group which generally opposed the adoption of the United States Constitution was the

1. Southern slaveowners.
2. debtor element.
3. commercial and trading interests along the Atlantic Seaboard.
4. well-to-do farmers who lived east of the Mississippi.

It would be quite appropriate as an alternative procedure to substitute for the above question one which asked why the debtor element opposed the adoption of the United States Constitution.

The usual reason for the inclusion of too much material in the alternatives is the fact that the item contains a series of problems rather than one central problem.

Another common error in writing suggested answers that makes them unnecessarily long is to include material in the alternatives which should have been included in the lead of the item. The following item illustrates this point:

A stone dropped from an airplane falls to the ground. This is a specific illustration of the

1. *general law that* all particles of matter attract one another.
2. *general law that* a force must be applied to a body to maintain it at rest.
3. *general law that* planets attract dense particles of matter.
4. *general law that* magnetized bodies attract other bodies.

In the example above the italicized material should have been included in the lead of the item.

It should be noted that there is no inconsistency between having brief alternatives and requiring the student to think.

4. Irrelevant cues should direct the examinee away from the right answer, if he is unable to solve the problem. They should never direct him towards the right answer.

There are several common ways in which the student is either led directly to the right answer or led away from a wrong answer.

Cues which lead the student to choose an answer without having a rational basis for his choice are called *specific determiners.* These cues should be used correctly in constructing tests. They are not things to be entirely avoided. If they are incorrectly used, they will lead the student who does not know the right answer to make a correct choice; but if they are properly used, they will lead the student who does not know the answer to choose one of the suggested wrong solutions. In any well-written item there are specific determiners operating, and the student who cannot solve the problem should tend to choose one of the incorrect alternatives.

Since there are certain specific determiners which should either be avoided or used with the greatest care, these will be considered in the following paragraphs.

A. If there is a close association through similarity of wording between the problem and the correct answer, it is likely that the student will be able to answer the question, even though he has inadequate background knowledge and understanding. The following item illustrates a specific determiner of this kind:

The basic unlearned element in the development of a conditioned *reflex* is a

1. *reflex.*
2. learned method of response.
3. glandular response.
4. memory trace.

In this item, the correct answer is a repetition of a key word in the lead and tends, therefore, to be highlighted.

Another example of the incorrect use of a specific determiner of this kind is given below:

What is the nature of a direct current according to the *electron* theory?

1. An exchange of positive and negative charges between each molecule and the next
2. A current of protons in one direction
3. A migration of neutrons and protons in one direction
4. A flow of *electrons* in one direction

Note how the validity of the item would be changed if the fourth and correct alternative were changed to read:

4. A flow of negatively charged particles in one direction

B. A mere association of sound between a key word in the lead and a key word in an alternative may result in the student selecting that alternative.

What is the meaning of the word *illicit*?

1. bad-tempered
2. impersonal
3. *ill*-mannered
4. elusive
5. *ill*egal

In the above item, both alternatives 3 and 5 have clang associations with the key word in the lead. While a clang association is used correctly in alternative 3, it should not be used in alternative 5, which is the correct answer. The

present writer suggests that the item might be improved by changing alternative 5 to "not permitted."

C. Alternatives that are drawn from a different domain than the one under consideration can be eliminated by the examinee without further understanding of the problem. Consider, for example, the following problem:

> Which American field commander of World War II most frequently embarrassed his superiors through making inopportune comments on political matters?
>
> 1. General George Patton
> 2. General Omar Bradley
> 3. General Mark Clark
> 4. General Bernard Montgomery

The last alternative in the item above is drawn from a different domain from that of the other three. In order to eliminate that alternative from the possible choices it is necessary for the examinee to know only that Montgomery was a general in a foreign army. In this item the fourth alternative should be replaced by the name of an American general.

D. As in the case of true-false items, absolute and all-inclusive terms such as *never, always, sole,* and *all* should not be included in wrong alternatives. They may, on occasion, be used in the correct answer.

Similarly, modifiers such as *sometimes, generally,* and *usually* should be used in both correct and incorrect alternatives, if they are used at all, since they may otherwise indicate the right answer. The correct use of specific determiners in multiple-choice items is the same as in the case of true-false items.

Examine the following item, which illustrates the correct

use of a specific determiner. (The fourth answer is the correct one.)

Under what conditions does Ohm's Law hold for a current in a conductor?

1. Only when the conductor is metallic
2. Only when the conductor is solid
3. Only when the conductor is either solid or liquid
4. Under all conditions

E. Another type of specific determiner which should be carefully avoided is one in which the lead of the item does not make a grammatically correct sentence in combination with each alternative. This occurs only when the lead is an incomplete sentence. The following item illustrates a cue of this kind:

Nausea and fainting spells which are conversion hysterias would be described by the behaviorists as

1. conditioned smooth-muscle responses.
2. reflexes.
3. instinctive behavior.
4. result of the extinction of a learned response.

In this item, it is evident that the fourth alternative was added as an afterthought without too much consideration for the lead which it was supposed to complete. The test-wise student knows that, in most cases, the alternatives which do not complete the lead grammatically are incorrect.

F. A common error of test writers is to include non-

functional alternatives because functional alternatives cannot be found. In a series of tests produced some years ago there appeared the following error:

> How does the gravitational force (G_1) exerted by the earth on a body on its surface compare with the gravitational force (G_2) exerted by the moon on a body of similar mass on its surface?
>
> 1. $G_1 > G_2$
> 2. $G_1 < G_2$
> 3. $G_1 = G_2$
> 4. None of the above

Since the first three alternatives include all possible cases, the last alternative is ridiculous and was included because the test writer could not think of anything else to include. This practice should never be followed. When such situations arise there are only two rational courses of action. One is to leave the item with only three alternatives. The other is to discard the item. If it is easy to replace the item with one of equal significance which has the desired number of functioning alternatives this should be done, since there are mechanical advantages in setting up items in some uniform style. In most cases there is little difficulty in developing items with a given number of alternatives, but where there is difficulty it is obviously much better to keep an item with too few alternatives than to discard it without replacing it.

G. Answers which are much longer or much shorter than the others should be avoided. They should never be right answers. Long and complex answers often stand out as right answers because they are more complete than the others. The following item illustrates the incorrect use of this type

of specific determiner. Note that the first and correct answer is much more elaborate than the others:

> In the United States it is possible that an individual may be a qualified voter in national elections in one state, but another individual in another state, possessing the same qualifications, may not be qualified to vote in national elections. This is possible because
>
> 1. the Constitution does not define voting qualifications for national elections, but leaves this to be decided by the state, subject to certain restrictions.
> 2. some states do not allow women to vote in national elections but others do.
> 3. many people fail to exercise their suffrage privileges in some states.
> 4. some states require voters to be natural-born citizens.

H. The answers to mathematical items are commonly given away by the fact that only one of the alternatives is *not* in numerical order. In order to prevent this defect from occurring, it is wise to arrange all numerical answers in order of size.

There is one exception to this rule, which certain test technicians observe. It is illustrated by the following set of alternatives to a mathematical problem:

(1) 0
(2) 1
(3) 2
(4) 3

In this case the second alternative is the correct one, but students who are aware of this fact may make the error of

marking the first alternative on their answer sheets because of the confusion between the answer and the number of the answer. In order to eliminate this possible difficulty some test technicians arrange the answers so that they correspond with the numbers of the alternatives in the following way:

(1) 1
(2) 2
(3) 3
(4) 0

THE CONTROL OF THE DIFFICULTY OF TEST QUESTIONS

In theory, a multiple-choice question can be changed in difficulty either by modifying the problem or by modifying the proposed answers. In practice, it is more satisfactory to control the difficulty of an item by controlling the difficulty of the problem. The reason for this is fairly obvious. Once a problem has been stated, the domain from which plausible alternative responses can be selected is limited. If an attempt is made to provide easier alternatives, it is probable that the correct solution will stand out very obviously from the others. The difficulty level of a test should be established by selecting those problems which the pupils should be able to solve if they have achieved the goals of the course.

Items are often made too easy unintentionally by the inclusion of alternatives which have only a remote relationship to the problem. The alternatives which fail to operate as distractors reduce the difficulty of the item and usually serve no useful purpose. It may be well to reemphasize

here this basic rule for constructing multiple-choice items, namely: *the decoys should distract.*

Those who are concerned primarily with theoretical problems connected with the making of tests may object to the viewpoint expressed in this section, which is derived entirely from practical experience with the construction of tests. There is nothing theoretically wrong with the idea of varying the difficulty of an item by changing the alternatives, but it has generally been found an impractical procedure because of the limited range of responses that can be included on any rational basis.

Rare words should never be used in a test question in order to add to the difficulty except when the purpose of the item is to test the student's knowledge of that particular word. They represent a poor method of adding to the difficulty.

Chapter Six

The Assembly, Administration, and Scoring of the Test

ONCE THE ITEMS have been prepared, it is necessary to check them against the original test plan. Under ideal conditions there should be one or more items corresponding to each cell in the test plan in which there is an entry. It is necessary to check the items against the test plan in order to insure that they are properly distributed and to determine, if necessary, the areas in which additional items need to be built. However, it has been previously noted that the behaviors represented by certain cells cannot be measured in paper-and-pencil situations. On the basis of this check, the teacher should select a group of test items for inclusion in the final test.

It would be rash for a teacher to gather together the items thus prepared and to reproduce the material as a test without making some further check on it. Material which is reproduced without further work is likely to include an unduly large number of ambiguous items, and items which have no correct answers. Consequently, it is most desirable to have the material worked over by another teacher, and the best method of obtaining this final check is to ask the

other teacher to work through the test, marking the correct answers on the answer sheet. The chances are that the answers thus recorded will not correspond in all cases with those which the writer of the test would have made. It should be noted that no test is so easy that this checking process can be eliminated. Ambiguous items appear in all tests as far down the age scale as language functions can be measured.

THE ARRANGEMENT OF THE ITEMS IN THE TEST

There is no formula available for determining the best method of arranging the items in a test. Each one of the methods commonly used has its own peculiar advantages and disadvantages, which will be discussed here.

1. Arrangement in Order of Difficulty. This is the most common method of arranging items in a test. The difficulty of the items is determined either on the basis of objective statistical information or by subjective judgment.[1] In both cases the true order of difficulty varies from one examinee to another. The major advantage of this method of arranging items in a test is that the student usually encounters the easier problems first and is not discouraged early in the test by encountering hard problems. One major disadvantage is that the items are not grouped either according to subject matter or according to the outcomes they attempt to measure. Some teachers feel that it is better to group together those problems that are similar so that the student will not have to switch rapidly from one domain of thought to another. A second major disadvantage of this arrangement is

[1] There is some evidence to show that teachers in a given subject-matter area can estimate the difficulty of items with considerable accuracy.

that the student may give up when he begins to encounter hard items, since he knows that those further on in the test are even harder. However, it is possible that some of the later items in the test might be easy for this particular student.

2. Arrangement in Cyclic Order of Difficulty. In order to avoid the second disadvantage mentioned above some authorities have advocated the arrangement of items in cycles of difficulty. The object of this procedure is to encourage the student to read every item in the test, for he knows that if he reads on far enough the items will become easy again, There is no doubt that this arrangement of the items has considerable merit, but it does not overcome the objection that it usually requires the student to switch rapidly from one subject-matter area to another.

3. Arrangement According to Subject-Matter Area. This system has the merit that it permits the student to think consistently about the problems of one subject-matter area before he passes on to the next. If it is used, it is desirable to combine it with the previous method and to arrange items in order of difficulty within each area.

4. Arrangement According to the Goals Measured. In some tests the publishers have grouped together those items that attempt to measure similar outcomes. For example, in some of the Cooperative Tests, items which measure knowledge of terms and concepts are grouped together, and those which measure more complex forms of understanding make up another group. The main advantage of this form of grouping is that it permits the teacher to see clearly the objectives that are being measured. The disadvantage lies in the fact that it does not usually permit the grouping together of problems from similar areas of subject matter.

The Arrangement of True-False Items

A special problem is presented by true-false items which must be arranged not only according to the factors previously discussed but also according to whether they are true or false. In general, true and false items should appear in random order,[2] but it is undesirable to have long runs of true or false items early in the test because such runs are easily remembered and the information may be passed on to those who are taking the examination at a later hour. There is no particular objection to long runs of true items or of false items towards the middle of the test.

Related to this problem is that of the fraction of the items in a true-false test which should be false. Common belief is that there should always be roughly an equal number of true and false items in a test, but this belief has no proper foundation. However, if a test is to be used on more than one occasion it is desirable to have approximately equal numbers of true and false items, for if there is a great excess of either true or false items the student who knows this fact before starting the examination gains considerable advantage.

THE TEST-ITEM FILE

The work involved in preparing good objective measures of achievement may make the teacher wonder whether the results are sufficiently important to justify it. Those who

[2] A slightly more sophisticated method than that of just mixing the true and false items is to use a table of logarithms for arranging them in random order. The procedure is a simple one. Start at the top of any column of common logarithms. Note whether the last digit is an even or odd number. If it is even, the first item in the test is true; if it is odd, the first item is false. The second number in the column is then examined. If the last digit is even, the second item is true; if it is odd, the item is false. and so forth.

have worked with objective tests believe that the total work involved in objective testing is no greater than that involved in using the traditional essay type of examination. In objective examinations most of the work is done before the test is administered. In essay examinations it is usual for the scoring process to involve a great amount of labor, while the test-construction process consumes relatively little of most teachers' time.[3]

In order for objective examinations to be reasonably economical in terms of the teacher's time, it is necessary that the materials prepared on each occasion be carefully conserved for future use. One good way of doing this is to file all test items on 5 by 8-inch cards, entering only one item on each card. It is also common practice to enter on the card statistical data concerning each item, if the item has been given to large groups.[4]

The cards should be classified in such a way in the item file that those needed for a particular occasion can be easily pulled. One convenient method is to classify the cards according to subject-matter areas and then to indicate on each by means of a number (number code) the objective which the particular item measures.

When an achievement test is to be prepared from the test file, a plan is first drawn up. Item cards are then pulled from the file to correspond with the entries in the blueprint. If there are no items in the file corresponding with certain entries in the test plan, then an attempt is made to prepare

[3] The present writer believes that well-constructed essay examinations require more labor to construct than objective tests. The common procedure of throwing together a number of questions a few hours before an examination is administered represents a most inadequate measuring technique.

[4] It is assumed throughout this manual that most of the tests developed by teachers will be given to relatively small groups and that little can be done to obtain cumulative statistical records on individual items. However, a brief discussion of this problem is found on pages 153–157.

new items to fill these vacancies in the evaluation procedure.

The test is typed directly from the assembled cards, which are refiled after the test has been given and the results of any item analyses have been entered on them.

In order that an item file may be used effectively, it is necessary to insure that all test copies given out to students be returned to the teacher at the end of the examination. This does not present a major problem where groups of 25 to 35 students are being examined in one room. Under such conditions it is a simple matter to number each examination copy, to pass out copies individually, and before the student begins work, to require him to hand back a 3 by 5-inch card on which he records both his name and the number of the examination copy he has received. This card is a receipt for the test booklet and places on the student the responsibility for returning that particular booklet.

While it is relatively easy to prevent test copies from being taken away from the examination room, it is much more difficult to prevent students from bringing away information concerning specific test questions. This becomes a serious problem in large universities where the same objective examination may be given to thirty or more sections in one day. Under such conditions it has sometimes happened that groups of students have organized themselves so that one member brings out the answers to the first ten questions, another member brings out the answers to the second ten questions, and other members are given other definite assignments. On some occasions the result has been that after the examinations were given during the 8-o'clock period, many of those who took them at subsequent hours were already trained on prepared keys. There is one simple method of overcoming this difficulty. It is to prepare two versions of the examination that are identical except for the fact that they have the items in reverse order. The two

forms are distributed at random to the group to be examined. This procedure not only solves the problem, but also eliminates any suspicion that cheating may have played a part in a student's score.

DIRECTIONS TO THE STUDENT

The directions to the student should contain statements concerning the following matters:

1. The purpose of the test.
2. The time allowed for answering questions and the speed at which the student should work.
3. The extent to which the student should guess or not guess when he is not sure of the answer.
4. Instructions concerning the way in which the student is to record his answers.

A sample set of directions is printed on page 133.

The Purpose of the Test

Little comment is necessary concerning this. The teacher should have no difficulty in interpreting the test to the student if it has been properly planned in the first place.

The Time Allowed

The first thing to note in connection with the time allowed for answering the questions is that, in general, there are relatively few cases in which speed is an important educational outcome. So long as speed is *not* an outcome, it should not be a factor in the student's performance on a test. Most tests of achievement should be power tests rather than speed tests, and enough time should be allowed for the examinee to do all he can do. This does not mean that the

DIRECTIONS

This is a test of your achievement in your first course in algebra which you are now completing.

There are 70 questions in the test. Answer the easier ones first and then return to the more difficult ones later. You may answer a question when you are not perfectly sure of the right answer, but avoid wild guessing. You will have 55 minutes for answering the test questions.

Five answers are suggested for each question. Select the answer which you think is right and then, on the separate answer sheet, blacken the space between the dotted lines under the small number corresponding to that answer.

All answers go on the answer sheet. Be sure to mark your answers with the special pencil provided [5] and make your marks heavy and black. Here is a sample question to show you how to mark your answers on the answer sheet:

Sample:

$$\frac{x}{2}+\frac{2}{x}=$$

1. x
2. $1\frac{1}{2}x$
3. $2x$
4. $2\frac{1}{2}x$
5. $4x$

Sample:

The first answer is the right one, so the space between the first pair of dotted lines has been darkened.

Scratch paper is provided for this test. Make no marks on the booklet, so that it can be used again with other students.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO

[5] These directions are for a test which is to be machine-scored. Therefore, a special pencil must be used.

examinees should be allowed to stay with the examination just as long as they want to sit, for there are always a few who will worry over the examination long after they have done all they can do on it. While time limits should be liberal, it is not good to waste the student's and the teacher's time by permitting the examinee to just sit and worry.

There are no rigid rules concerning how many items an individual can complete during a given time. The number of items that can be finished depends on the level of ability of the examinees, their age, the difficulty of the examination, and the kinds of functions that are measured by the test. However, at the high-school level, students may be expected to complete at least 150 vocabulary items per hour. Reading-comprehension items usually require at least a minute each, and other items that call for careful thought may require even more time. Unfortunately, a common tendency has been to include too many items in tests because of the need for obtaining high reliability in a brief space of time. Many comprehensive-examination programs make this basic error; while students should be required to work as fast as they can, it is usually desirable to provide enough time to permit even the slowest student to show what he can do.

To a considerable extent the student will adjust his speed to the amount of work expected of him. The fact that five per cent of the students do not finish a test does not mean that the test is too long, and in some cases it may actually be lengthened without increasing the percentage of those who would still like to take more time than is given.

It is a good plan to advise students in taking objective tests to go over the test, doing the easier items first, and then to go back to the more difficult items later. In this way the student gets at an early stage a good idea of the amount of work to be done in the allotted time.

Guessing

In the matter of whether the student should guess when he does not know the answer, there are two procedures which can be rationally justified. They are:

1. The examiner may instruct the students to attempt every test question, regardless of whether they know the answer. If this is done, there is no value in subtracting some fraction of the wrong answers from the right answers, since to do so will not alter the order in which the examinees are placed.

2. Alternatively, the examiner may instruct the examinees to answer questions when they have some idea of the right answer, but to avoid wild guessing. In this case a correction for guessing should be applied in order that the over-confident student may not gain undue advantage. In general, the number of wrong answers should be divided by one less than the number of alternatives in the questions and the quotient should be subtracted from the number of right answers. Thus, if an examinee is given 100 true-false questions and marks 80 right and 16 wrong, his corrected score would be $80 - \frac{16}{2-1} = 64$, since a true-false question is a two-answer question. If an examinee is given 90 questions with 4 answers from which to choose in each question, and if he marks 50 right and 18 wrong, his corrected score would be $50 - \frac{18}{4-1} = 44$.[6] The correction

[6] This formula assumes that when the student does not know the answer he will select an alternative at random. This, of course, may not be the case, since the items have been constructed with the hope that the student who does not know the correct answer will prefer to choose one of the incorrect alternatives. In so far as the test has been successfully constructed along these lines, there should be a special correction not only for each test but for each situation in which it is used. The correction should be such that the corrected score correlates to the highest possible extent with

is such that, by guessing throughout the examination, the examinee may be expected to get a score of zero.

In practice, most justifiable marking schemes reduce themselves to one or the other of these two alternative systems. Other marking plans should be avoided in most cases since they tend either to favor or to penalize the overconfident student.

The Recording of Answers

The teacher can save much labor by requiring the student to record his answers on a separate answer sheet rather than on the sheets containing the test questions. The advantages of this procedure are twofold. First, it permits the re-use of the test booklets and second, it often reduces the scoring time by as much as 90 per cent. The design of answer sheets and scoring procedures will be considered in the next paragraph, since they are closely related to one another.

METHODS OF SCORING TESTS

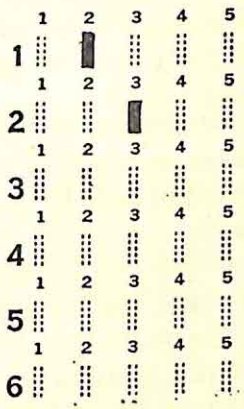
The development of test-scoring machines has raised the hope in the minds of teachers that soon many schools will have mechanical means of scoring examinations. Since the advantages of test-scoring machines are much better known and appreciated than their limitations, mainly the limitations of machine methods of scoring tests will be discussed here.

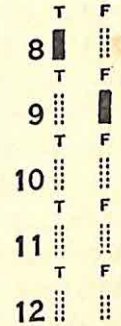
The International Business Machines Corporation scoring machine is a device which will identify correctly placed or incorrectly placed graphite marks on a sheet of paper. The machine will count the number of graphite marks that

some appropriate criterion. However, it is usually quite impossible to develop such special formulae since no satisfactory independent criterion can be found. Consequently, the above formula is used in spite of the fact that it is based on questionable assumptions.

are correctly placed and the number that are incorrectly placed and will give these totals separately. The machine will also subtract any particular desired fraction of the wrong marks from the right marks and will perform this scoring operation at any rate up to 600 answer sheets per hour.

In order to use the machine a special answer sheet must be used. These special answer sheets are designed in a great variety of forms. Some are for true-false questions, others for multiple-choice questions, and still others provide space for both types of questions. Sections from two kinds of machine-scored answer sheets are shown below and the responses to two items have been marked on each sample:





The machine scoring process is satisfactory only if certain conditions are fulfilled: (1) the student must use a special graphite pencil; (2) he must make heavy marks on the paper and must not make any stray marks; (3) he must use a special answer sheet which costs between ½ cent and 1 cent depending on the kind selected; and (4) scoring must be undertaken on a dry day when the machine is operating properly, since the machine cannot be relied upon to work effectively on humid days.

In order to insure that the first two of these four conditions have been fulfilled, marked answer sheets must be carefully inspected before they are scored by machine, all stray marks must be erased, and lightly marked answer sheets must either be separated for hand scoring or be re-marked with heavy pencil. The inspection of the answer sheets is a more laborious process than the scoring procedure itself. When the answer sheets have been scanned for defects and are ready for scoring, the machine must be set up and checked for the particular test which it is desired to score.

When the papers have been scored, it is desirable to have them rescored. Some large organizations rescore by hand, others rescore on a different machine. Hand rescoring is preferable because it is likely to catch errors made by the machine in the units digit.[7] The human scorer often makes errors of ten, while the machine tends to make errors of unity and few errors of larger magnitude.

There are certain situations in which scoring machines are invaluable. These situations occur in organizations such as universities, large school systems, and state Civil Service commissions, where large numbers of individuals are given uniform batteries of tests. For example, if a school system of considerable size were to administer the same series of aptitude tests to all those in a given grade, the scoring of the tests would be undertaken best in a central scoring unit equipped with one or more scoring machines. However, the whole process of obtaining special pencils, scanning the answer sheets, setting up the machine for scoring, and rechecking is so elaborate that it is practical only where large numbers are to be examined, and where there is a central organization to undertake the work. While the machine

[7] Under ideal conditions scoring and rescoring would be undertaken by two methods which produce uncorrelated errors. Scoring and rescoring by the same method are likely to produce correlated errors which will be undetected by the procedure.

age has come to education, it results in improvement only in cases where the mass production of test results is involved.

Much can be done to improve hand-scoring techniques. For example, whenever objective tests are given, the answers should be recorded on a separate answer sheet in order to facilitate the scoring process. If the answers to objective-test questions are recorded directly in the test booklet, scoring is a slow and tedious process. Many individuals use the answer sheets of the International Business Machines Corporation for all objective tests and score them by hand, but a mimeographed answer sheet is just as satisfactory and costs a great deal less. An answer sheet of the following type will be found to be satisfactory for most tests consisting of best-answer type questions:

1. a b c d e
2. a b c d e
3. a b c d e
4. a b d c e

Similar answer sheets may be used for recording responses to true-false items. International Business Machines Corporation sells an answer sheet for exclusive use with true-false items and another which permits both true-false and multiple-choice answers to be recorded on the same sheet. Home-made mimeographed answer sheets can also be used with true-false items and would look like the following sample:

1. T F
2. T F
3. T F
4. T F
5. T F

In no case should T F be typed before every question on the question sheet, because a question sheet so arranged is both hard to score and troublesome to most students, who would need to be left-handed to mark the answers without inconvenience.

Hand scoring of this type of test is best undertaken by means of a stencil key made by punching holes in a blank answer sheet. Many teachers have had difficulty in obtaining a punch which will make holes in the middle of an ordinary sheet of paper but this difficulty has been overcome by a special punch developed by International Business Machines Corporation for making stencil keys. This punch sells for $5.00 and is by far the most practical device for the preparation of keys. It will punch a hole in the middle of an 8½ by 11-inch sheet. It is suggested that students be asked to indicate their answers by means of a cross on the mimeographed answer sheet. A cross will easily be seen through a hole in a stencil, while a circle cannot be seen through the hole.

If wrong answers are counted, and if there is only one right answer to each question, the following procedure will insure accuracy of scoring:

1. The stencil is placed over the answer sheet and each correctly marked right answer is marked with a red pencil and counted. The number of right answers is then noted on the answer sheet.

2. The stencil is removed from the answer sheet and the wrong answers are then counted by means of a stencil which is punched out for the wrong answers.

3. The total number of marks or crosses on the answer sheet is counted and this count should check against the number right plus the number wrong. Time may be saved by asking the examinees to count the number of items

they have attempted and to record this number on the answer sheet. Alternatively, a count may be made of the questions *not* answered on the answer sheet. The number right plus the number wrong, plus the number omitted should be the same for all answer sheets provided that not more than one response is indicated to each question.

THE REPRODUCTION OF TESTS

All tests should be set up for reproduction and reproduced in such a way that they are easily read. Regardless of whether tests are reproduced by photo-offset, mimeograph, Ditto, or letterpress, it is preferable to set up multiple-choice items in two columns on an 8½ by 11-inch sheet rather than to print them right across the page. The two-column form is more economical of space than the single-column form.

The alternatives for the items should be printed one below the other, just as they are printed in this book. This layout facilitates the reading of the alternatives, and if the student knows the answer to the problem he can easily find it among the alternatives. Never split an item at the bottom of a column. The item and the alternatives should form a single unit. If there is insufficient space at the bottom of a column for a complete item, the item should be started at the top of the next column.

Most tests developed by teachers are likely to be mimeographed because the equipment for doing this is usually available, but if diagrams are used in the test it is easier to reproduce it by the Ditto process. Larger quantities of tests can be conveniently reproduced by photo-offset. This is accomplished by photographing the typed pages provided by the teacher. One of the major advantages of photo-offset is that it eliminates any type-setting process and it is thus un-

necessary to wait for galley proofs and page proofs, since all proofreading is done on the single typed copy provided for the printer. Photo-offset reproduction is generally cheaper than regular printing for quantities up to 10,000.

Chapter Seven

The Significance of Test Scores

A RAW SCORE on an achievement test has no significance unless additional data for interpreting it are available. These supplementary data may be of various kinds.

1. The commonest form of supplementary information is supplied by a table of norms which indicates the performance of some specified group on the test. Norms are given in percentiles and in various kinds of standard scores. From such a table, it is possible to determine where a child or adult stands with respect to a specified group. It is not possible without further information to determine what goals have been achieved in the person who takes a test. Tests thus standardized are prepared in such a way that some members of the standardization group will obtain high scores and others low scores. Unfortunately, in the past many teachers have been satisfied with knowing only a pupil's relative standing in a group. Many of the evils that are attributed to the assignment of grades result from the use of this system for interpreting test scores. It is obvious that according to this system some must fail because others pass. While failure is sometimes an excellent, though pain-

ful educational experience, the kind of consistent failure that results from this type of grading system is likely to be demoralizing.

2. Scores may also be interpreted in terms of the extent to which the outcomes of teaching have been achieved. The supplementary information needed in this case is the test plan showing the relationship of the test to the goals. Under such conditions, the primary objective is *not* to obtain a distribution of scores, for all students may have achieved or failed to achieve the desired outcomes.

While the second approach to the interpretation of scores is more acceptable than the first, there are basic difficulties inherent in its application. The main problem arises from the fact that when teachers are asked to indicate in advance the minimum scores they will accept as evidence of the achievement of their objectives, they are likely to designate achievements which only a few students ever attain. This is largely a result of the fact that not too much is known as yet concerning the outcomes which can be achieved at various levels of ability. Teacher judgments will become much more realistic when accurate knowledge concerning the student's achievement potential is available.

BASIC CONSIDERATIONS IN DEVELOPING A GRADING SYSTEM

Any system of assigning grades inevitably includes many unsatisfactory elements since the scale does not, and possibly cannot, meet the criteria of a satisfactory measuring device. The chief deficiencies of a grade scale, namely that it does not have an absolute zero nor a well-defined unit of measurement, cannot be remedied at the present time. However, something can be done to clarify what a given grade means

and to establish some uniformity of significance of grades throughout a school or throughout a school system. The following suggestions are made with the recognition that, although they may improve the present grading system, they do not overcome certain basic difficulties in the measurement of achievement.

A grade system is a measuring scale, but before it can be used as such, it is necessary to know what it is measuring. At the present time, student grades are used to measure at least three different things. The grades given by one teacher may indicate the extent to which the goals of the course have been achieved in the students. The grades given by a second teacher may indicate the relative amount of progress made by a student. The grades given by a third may indicate the relative standing of the students (grading on the curve). According to the system of the second teacher, a student who starts a course with a poor background and finishes with average proficiency may deserve a better grade than one who starts out with a good background and finishes the course with a high level of achievement. A school should determine what is to be measured by grades so that those who use the cumulative records may know how to interpret them. Most cumulative records cannot be interpreted because the grades of different teachers mean different things.

The present writer believes that the only grading system which has a sound rational basis and provides interpretable grades on the cumulative record is the one which indicates the extent to which the goals of teaching have been achieved. In order that grades in such a system may have meaning it is necessary that they be based on some scheme such as the following, which, it may be noted, applies only when the achievement of one set of goals is necessary before more advanced goals can be achieved. Not all goals are structured in a hierarchy.

GRADE	SIGNIFICANCE OF GRADE
A	All major and minor goals have been achieved and the achievement level is considerably above the minimum required for doing more advanced work in the same field.
B	All major goals have been achieved, but the student has failed to achieve some of the less important goals. However, the student has progressed to the point where the goals of work at the next level can be easily achieved in him.
C	All major goals have been achieved, but many of the minor goals have not been achieved. In this grade range the minimum level of proficiency represents a person who has achieved the major goals to the minimum acceptable extent and hence who has the minimum amount of preparation necessary for taking more advanced work in the same field, but without any major handicap of inadequacy in his background.
D	A few of the major goals may have been achieved but the student's achievement is so limited that he is not well prepared to work at a more advanced level in the same field.
E	None of the major goals have been achieved.

Students within each of the grade groups above except the last may be divided into three categories (such as A+, A, A—) so that grading may be undertaken on a 13-point scale. It should be noted, however, that this scheme is only one of many that could be developed. The grading system

necessarily reflects the educational philosophy upon which it is based.

The fraction of the students obtaining grades of A in a given course will depend largely on the extent to which the goals are easy or hard to achieve and probably to a much smaller degree, sad to say, on the skill of the teacher in achieving those goals in the students. The first of these factors is subject to administrative control, the second factor is not. Consequently, it might be administratively desirable to request teachers to set goals of a kind that will make it possible for all students in the course to achieve the major goals to a minimum extent. Unfortunately, there will be a considerable fraction of the students who, through lack of motivation and other causes, will fall in the lower categories at the end of the semester even when this policy is adopted. Some will obtain low grades because of the lack of reliable guidance procedures, some because of lack of motivation, some because of factors disturbing their personal life.

The system proposed is recognized as being far from completely satisfactory, but until the outcomes of education can be measured with some precision and until the goals of education can be defined with equal precision, the best grading system will leave much to be desired. However, it is believed that the system outlined above will bring some order into a very confused situation and eliminate many of the evils of grading which result from this confusion.

If grades are to measure the extent to which the goals have been achieved in the student, then the average grade given to students in a course will indicate the degree of ease or difficulty with which the goals can be achieved. If all students are given grades of A in a course, the indications are that the goals are fairly easily achieved in the student group. The situation suggests, not that the grading system is too lenient, but that the educational goals are possibly,

but not necessarily, too readily achieved. If all the students satisfy all the requirements of a course, they are entitled to A's, and any pressure to reduce the number of A's is manifestly unjust. In such a situation, a strong case can be made out for adding goals to the course that are more difficult to achieve. The purpose of this would be not to increase the number of failures but to raise the goals to be achieved to a point where all might have to work efficiently to attain these goals. Obviously an educational program should be such that it produces a maximum rather than a minimum change in students in a given time.

THE VALIDITY OF ACHIEVEMENT TESTS

This entire volume has been concerned fundamentally with the validity of achievement tests, though the term *validity* has been hardly used at all. It is probably unnecessary to point out to the reader that the *validity* of a test is the extent to which it measures what it is supposed to measure. The most difficult problem in the construction of achievement tests is to devise methods of determining the validity of the tests; in general, there are two main approaches to the solution of this problem.

One approach is to find some independent criterion of the achievement of educational outcomes and to compare measures derived from the evaluation instrument with those derived from the independent criterion. This method works well in appraising the validity of tests of achievement in skilled trades such as that of the machinist or the carpenter. In such cases it is possible to identify groups of men in whom the objectives of training have been achieved and those who have acquired relatively little skill. If the achievement test discriminates between these two groups then the

discrimination provides some evidence of validity. This situation is clearly in contrast with that found in measuring the outcomes of an academic high school curriculum, where it is usually quite impossible to identify with certainty groups in whom the outcomes of instruction have been achieved. In this situation it is practically impossible to obtain a criterion of validity of an achievement test which involves an independent measure of achievement. Under such conditions the following approach to the determination of the validity of an achievement test must be used.

In this alternative approach the test is made valid by definition. That is to say, behaviors are listed which will be accepted as evidence of the achievement of the desired outcomes and problem situations are developed in which the student will manifest the desired responses if he has acquired them. For example, if it is desired to develop the ability to multiply two-digit numbers, then a test which requires the student to multiply two-digit numbers is necessarily valid. If it is desired to develop in the student a knowledge of the vocabulary of the social sciences, then a test-problem situation which can be solved *only* through having a knowledge of such vocabulary is necessarily valid by definition.

While this second approach is simple in conception, there are certain difficulties inherent in it. It assumes that there is identity of the behaviors that are to be achieved and the behaviors that test situations are designed to elicit. In actual practice there are relatively few situations in which complete identity exists. Usually, in a well-designed test, the test situations are similar to, but not identical with, the problems which the student has been trained to solve. For example, in a course in consumer mathematics the student may have been taught the basis for comparing two brands of canned vegetables and the teacher may have attempted to develop a willingness to compare brands in terms of the

value obtained for the money spent. The extent to which such objectives have been achieved can be determined only by observing the student in an unrehearsed buying situation in which he does not know that he is observed. The corresponding paper-and-pencil situation omits at least one important element and introduces an irrelevant but influential element, namely, a knowledge on the part of the student that he is being observed. If it were possible to show that the behaviors in the unrehearsed buying situation were closely correlated with those in the paper-and-pencil situation, evidence would have been provided of the validity of the paper-and-pencil test. Usually such evidence is not available, but every effort should be made to obtain it.

The second approach also assumes that evaluative criteria can be specified with such definiteness that they can be matched with behaviors shown by students in their responses to tests. Such is not the case since methods of specifying evaluative criteria are still rather crude.

It is quite unfortunate that one of the commonest procedures for "validating" achievement-test items seems itself to be largely lacking in validity. The procedure is simply that of determining the extent to which each item becomes progressively easier for students from each grade to the subsequent one. Those items that show this change in difficulty are assumed to be valid for inclusion in the test. It need hardly be pointed out that this procedure is wholly fallacious, for it is quite evident that growth may occur without important objectives being achieved. Unfortunately the most widely used achievement-test batteries for the elementary school have been validated in this way. The fact seems to be that the proper validation of achievement tests must involve a comparison of the behaviors elicited by the test-problem situation and behaviors which are to be accepted as evidence of the achievement of the desired outcome.

THE RELIABILITY OF TEACHER-MADE TESTS

Manuals supplied with commercial achievement tests usually present data on the reliability of the tests, that is to say on the consistency with which the tests measure whatever they do measure.[1] The argument is that if tests scores are to have any meaning at all, they must be based on reliable tests. Reliability is a necessary but not a sufficient condition for meaningful measurement. It is discussed at length in manuals on achievement tests because data on the validity of the tests is usually lacking.

It is usual for test writers to attempt to build achievement tests in such a way that the range of scores found when the test is administered corresponds to the total possible range of scores. This practice is designed to insure that the test will have reliability, but it disregards the fact that if all the goals have been achieved in all the individuals tested the range of scores should be zero.

The reliability of an achievement test should be determined with groups which include individuals ranging from those in whom none of the objectives have been achieved to those in whom all of the objectives have been achieved. Unfortunately most groups who have been through some kind of educational program are fairly near to meeting this condition. Unless one can assume that the condition stated at the beginning of this paragraph actually exists the calculated reliability of an achievement test cannot be interpreted. In the remainder of this section it is assumed that reliability is measured under this condition.

[1] This statement of the concept of reliability greatly oversimplifies the problem. However, it would not be profitable in this volume to discuss some of the more elaborate concepts of reliability which have been advanced during the last few years.

The reliability of tests is usually measured in terms of a correlation coefficient. A perfectly reliable test would have a reliability of 1.00 while a completely unreliable test would have a reliability of 0.00. Most commercial achievement tests will have reliabilities of 0.9 or greater.

The experience of the present writer leads him to the conclusion that most teacher-made objective examinations have rather low reliabilities for two main reasons: (1) They are likely to be too short to give adequate reliability. They are short, not because there is insufficient time in a class period to give a test of adequate length, but because the preparation of a sufficient number of objective-test questions requires a considerable amount of the teacher's time. This difficulty can be remedied only if the teacher keeps a cumulative file of test questions from which to build examinations. (2) There is a marked tendency for objective tests made by teachers to be too difficult. The result of this is that scores are bunched at the lower end of the distribution and there is a resulting loss in reliability. It is fairly obvious that if scores are bunched together near the score which a student might be expected to achieve on a chance basis, then differences between these scores have very little significance.

From the discussion above it is fairly evident that, in order to have reliable classroom examinations, teachers must (1) include a reasonably large number of items in the tests, and (2) include a sufficient number of easy items to prevent the bunching of scores at the lower end of the distribution. Unfortunately there is no rule-of-thumb method of determining whether a test has sufficient length to give it reasonable reliability. However, it will rarely happen that a teacher-made test which has sixty multiple-choice items will have a reliability above 0.8. In the experience of the present writer, it is necessary in most fields outside of mathematics to include from 80 to 100 items in a test in order that

it may have reliability of at least 0.8 the first time it is given. Ideally, a test should first be tried out experimentally, and the difficulties should be adjusted by adding easier or harder items and by removing the items which do not contribute to the reliability of the test. However, this procedure is impractical in most school situations, and for that matter in numerous situations outside of school, since many examinations cannot be tried out before they are actually used for testing purposes.

HOW THE TEACHER MAY USE ITEM ANALYSES

Teachers are usually most interested in the total scores that pupils make on a test. Sometimes they are concerned with part scores that may indicate the extent to which each of several objectives has been achieved. Only rarely are teachers concerned with the responses of pupils to individual items on a test, but when they are so concerned an item analysis will provide useful information. For example, if the teacher should give a test consisting of 50 multiple-choice questions, she might wish to find out which questions were the hardest, which questions were the easiest, which questions the pupils misunderstood, and what common misconceptions were held by the students. Some information can be obtained about these problems if the number of pupils who selected each alternative is counted. This is a straightforward and simple task if there are only between 30 and 40 students in the class, and the information obtained is worth the time devoted to procuring it. However, from such an item analysis it is not possible to make generalizations concerning the difficulty of the same items for other groups. Neither is it possible to generalize to any extent

about the misconceptions which other groups of students may hold.

The kind of item analysis discussed above cannot be used for revising the test materials, since it is based on limited data. It is carried out entirely for the purpose of determining what has happened in a particular situation. In most cases, an item analysis of this kind is all the item analysis that the teacher can profitably make because of the small size of the group tested. However, there are times when objective achievement tests are given to large groups—as when an examination is built and given on a departmental basis. In such cases it is desirable to make an item analysis not only for the purposes described above, but also for selecting those items which may be most profitably used on subsequent occasions.

When an item analysis is made for the purpose of selecting the best items for future use, it is necessary to have some criterion to use in the selection process. Under ideal conditions, it would be possible to obtain a group in which the particular outcome measured by the item had not been achieved and a group in which it had been achieved. The perfect test item would be one that would be passed by all in one group and failed by all in the other. Poor test items would be marked correctly about as frequently by one group as by the other.

Unfortunately, there are few instances in which it is possible to select items on the above basis, since it is hardly ever possible to identify with certainty groups in whom a particular educational outcome has been achieved. Consequently, other procedures are used, the most common of which is based on the assumption that the score on the test as a whole is a more valid measure of what it is desired to measure than the score derived from each individual item. On that basis it is argued that the correlation between the

performance on an individual item and performance on the test as a whole will indicate the extent to which an item is valid.

It is obvious that the latter argument is based on questionable assumptions, and in recognition of this fact it has become customary to refer to the correlations between item scores and total tests scores, not as coefficients of validity, but as coefficients of internal consistency. On this basis, the primary purpose of an item analysis is to select items which contribute most to the internal consistency of the test.

Several short-cut procedures have been devised for estimating the correlation between the performance of the students on each item and their performance on the test as a whole. A procedure which has been widely followed requires the following steps: [2]

1. A random sample of about 400 answer sheets is taken. A very convenient number of answer sheets is 370. However, it is not desirable to work with appreciably smaller numbers.

2. The answer sheets are arranged in order according to the size of the score on the test as a whole. This operation results in a pile of sheets with the paper having the largest score on the top and the one with the smallest on the bottom.

3. The top 27 per cent and the bottom 27 per cent of the papers are removed from the pile. If 370 papers have been used, the upper and lower 27 per cent will each consist of 100 answer sheets.

4. Each group of 100 answer sheets is then treated sep-

[2] A procedure which is more complex but in many ways more satisfactory than the one described can be found in the following reference: Davis, Frederic B., *Item Analysis Data*, Harvard Education Papers No. 2, 1946, pp. 42 + v. However, it should be noted that there are many well-founded theoretical objections to both of these systems. At best they provide a very rough estimate of what needs to be known about the homogeneity of the item with the rest of the test.

arately. First one group and then the other is taken, and the number of students who chose each alternative in each item is counted. This procedure will provide data of the following type for each item:

	Percentage in Lower Group Choosing Each Alternative	*Percentage in Upper Group Choosing Each Alternative*
Alternative A	23	8
Alternative B (right answer)	32	75
Alternative C	40	17
Alternative D	5	0

These data show that more persons in the upper group than in the lower group were able to answer the item correctly and that each of the wrong alternatives was chosen more frequently by the lower group than by the upper group. The data also indicate that the ability to select the right answer to the item is associated with the ability to obtain a high score on the test and that the item is not detracting from the reliability of the test as a whole.

It is possible from these data to estimate the correlation between the performance on the single item and the performance on the test as a whole. This can be done by referring to a chart prepared specially for this purpose.[3]

From the data given above it is also possible to estimate the *difficulty coefficient* of each item, that is to say, the percentage of the total group tested that answered the item correctly. This is done by finding the average number in

[3] Thorndike, Robert L. *Personnel Selection*, John Wiley and Sons, Inc., New York, 1949.

the upper and lower groups combined who answered the item correctly. In the case of the above item the difficulty coefficient would be $\frac{32 + 75}{2}$, or 53.5. It should be noted that this figure is an estimate and is not necessarily equal to the true value, which would have to be calculated from the total distribution. It is evident that the numerical value of the difficulty coefficient of the item derived from the tails of the distribution may differ considerably from the value derived from the total distribution.

Appendix

Objective Methods of Scoring Free-Answer Examinations

FREE-ANSWER QUESTIONS discussed in the main body of this text have been limited to those in which the response is confined to one or two words. It was pointed out that the general procedures for evaluating the outcomes of teaching discussed in this volume apply regardless of whether a free-answer or a controlled-answer type of examination is used. In each case, the initial procedure involves the identification of behaviors which are to be accepted as evidence of the achievement of the desired outcomes. These behaviors are then observed to occur or not to occur in test situations.

In the best-answer or true-false type of test it is easy to determine whether a given behavior occurs or does not occur, but in the free-answer type of test, difficulty is encountered because of the great variety of responses given by students. In most cases, few of the responses will be identical with the behaviors used to define the outcomes but many responses will approximate these behaviors sufficiently closely to be accepted as evidence of the desired achievement. The fact that the scorer's judgment determines whether a given response does or does not receive a score

results in considerable unreliability in the scoring process. This problem becomes progressively more acute as the free-answer test is extended from one requiring a single word as a response to one requiring a response of the essay type.

THE ESSAY OR FREE-ANSWER TEST AS AN OBJECTIVE EXAMINATION

The expression "essay examination" is commonly used by teachers in a loose sense to cover a multitude of free-response examinations. It is not usually restricted to the type of examination which involves the appraisal of a literary composition, but is used to cover all types of examinations except objective tests. It is used here in this broad sense to cover any examination in which the student composes his own response in contrast to the examination in which he selects one of several responses provided.

The problem of preparing and scoring essay questions is discussed in this appendix rather than in the main body of the text because in the essay examination the chief difficulties lie in the scoring procedure rather than in the preparation of the questions. With most types of objective questions, on the other hand, the main difficulty lies in the preparation of the test, while the scoring procedure is a relatively simple mechanical task.

Advice on the use of essay tests cannot be given with the same certainty as advice on the common forms of objective questions discussed in the rest of the book. Investigations undertaken many years ago showed the unsatisfactory nature of essay tests as they are commonly used by teachers and on that basis many people assumed that the essay was necessarily a poor evaluation instrument. The result has been that a prolific amount of work has been done on the

development of the multiple-choice type of test but practically nothing has been done to improve the essay as a measuring instrument.

Under certain conditions essay tests may be considered as objective examinations; that is to say, in some cases the rules for scoring them may be so precisely formulated that the correct score can be determined accurately by independent observers. Consideration will be given here only to those essay questions that can be scored in this way.

OBJECTIVE METHODS OF SCORING FREE-ANSWER OR ESSAY TESTS

1. Essay Questions Measuring Knowledge. Essay examinations may be used to determine the student's knowledge of facts. The following questions are fairly typical of those commonly used to measure this outcome:

What were the main functions of the W.P.A. (Works Progress Administration)?

How are fertilizers prepared from the air?

How is meat distributed from the farmer to the consumer?

Describe the organization of your city government.

Describe an experiment to demonstrate Archimedes' principle.

How are Presidential candidates selected prior to election day?

The common practice in scoring the responses is to read the essay and then to assign a mark on the basis of some gen-

eral impression. This method is most unsatisfactory, for if two people score the same papers on this basis there are likely to be considerable discrepancies between the scores they assign.

The following procedures for scoring essay examinations are usually considered to provide greater reliability than the method of assigning scores on the basis of a general impression. There is also some evidence to warrant the belief that they produce more valid scores.

Procedure A. (1) A model answer to the question is written by the teacher. (2) The model answer is dissected into as many separate points as is feasible and one or more marks are assigned to each point. The number of marks assigned to each point depends on the teacher's judgment of the importance of the particular point. (3) Each point included in the answer of each student is scored according to this system and the marks thus assigned are added up for each answer.

Procedure B. (1) Relevant points are listed from those given in the essays of the students. In order to do this, it is not usually necessary to go through all of the students' papers at one time. It is generally sufficient to list points from the first 10 or 15 in order to make up the master list of relevant points which are to receive scores. Additional points may be added at a later time. (2) A given number of marks is assigned to each point as in Procedure A. (3) Each student's essay is scored according to the master list of points and in this process additional points may be added to the master list. The marks for each essay are then added up.

Whichever procedure is used the task is a tedious one and where classes of 30 or more students are involved it becomes quite impractical for the teacher who can devote only a limited amount of time to systematic evaluation procedures.

Of course, there are cases in which an essay examination may be scored more validly on the basis of the general impression created than by the use of a detailed scoring system. An example would be an English-composition test on an advanced level.

2. Essay Questions Measuring the Application of Principles. Free-answer questions are commonly used in fields involving quantitative methods for the purpose of appraising the student's ability to apply mathematically stated principles to the solution of problems.

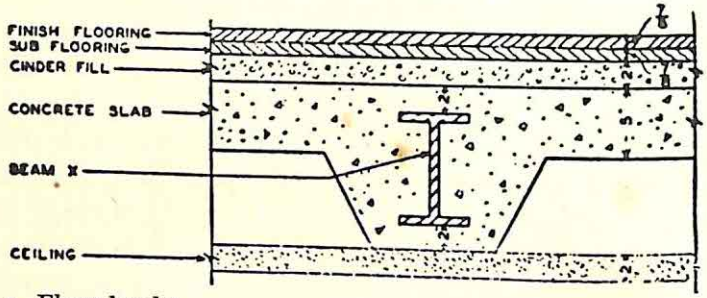
One common way of scoring the answers to such problems is to appraise them entirely in terms of whether the free answer was or was not correctly calculated. From the point of view of the examiner this procedure is by far the simplest one, but it fails to extract as much information as can be extracted from the test situation, for the student who makes a slip only in the last stage of solving the problem is penalized as heavily as the student who is in error throughout.

When the teacher recognizes that it is desirable to give partial credit for the partial solution of a problem, it is usual for him to appraise subjectively the seriousness of the errors and to mark accordingly. The subjective element in this process is a major source of unreliability and most experts recommend that it be eliminated by the same procedures used in evaluating the type of essay questions previously discussed. A check list is prepared by which marks are assigned to the solutions presented to the problem.

The following two problems illustrate the use of a check list for scoring numerical free-answer problems involving the application of scientific principles. It should be noted that the check list is not complete, but includes only selected items for illustrative purposes.

Problem 1.[1] The student is asked to determine the required size of beam X shown in the figure. The following data are given:

1. Floor-to-floor height in the structure, 12′ 8″
2. Ceiling height, 11′ 0″
3. Floor construction as shown in figure:



4. Floor loads:

Live load	100 lb. per sq. ft.
Movable partitions	18 lb. per sq. ft.
Finish flooring	3 lb. per sq. ft.
Sub-flooring	3 lb. per sq. ft.
2-inch cinder fill	16 lb. per sq. ft.
5-inch concrete slab	60 lb. per sq. ft.
Total	200 lb. per sq. ft.

5. Weight of beam plus fireproofing is assumed to be 120 lb. per linear foot.
6. Beams are placed 8 feet on centers.

Scoring Check List for the Examiner	*Yes*	*No*
1. Was the load of the floor per linear foot of beam determined to be 1,600 lb.?		
2. Was the weight of beam, plus fireproofing, used to determine total load per linear foot acting on one beam?		

[1] In William A. Brownell *et al.*, *The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, pp. 296–297). The University of Chicago Press, 1946.

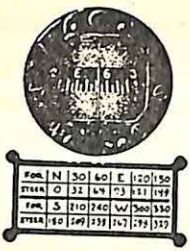
3. Was the total load per linear foot acting on one beam found to be 1,720 lb.? .....
4. Was the total load on one beam found to be 30,960 lb.? .....
5. Was the maximum bending moment found to be 835,900 lb.? .....
6. Was the required section modulus found to be 46.4 inches cubed? .....
7. Was the maximum allowable height of beam found to be 10¼ inches? .....
8. Was a 10-inch wide flange, 45-lb. I-beam selected from the tables of the Steel Construction Handbook of the American Institute of Steel Construction? .....
9. Was the actual section modulus of the beam determined from the tables to be 49.1 inches cubed? .....

Problem 2.[2] A paratroop transport with air speed of 190 miles per hour is to fly from Leuchars, Fife, Scotland, to Bergen, Norway. The following data are given:

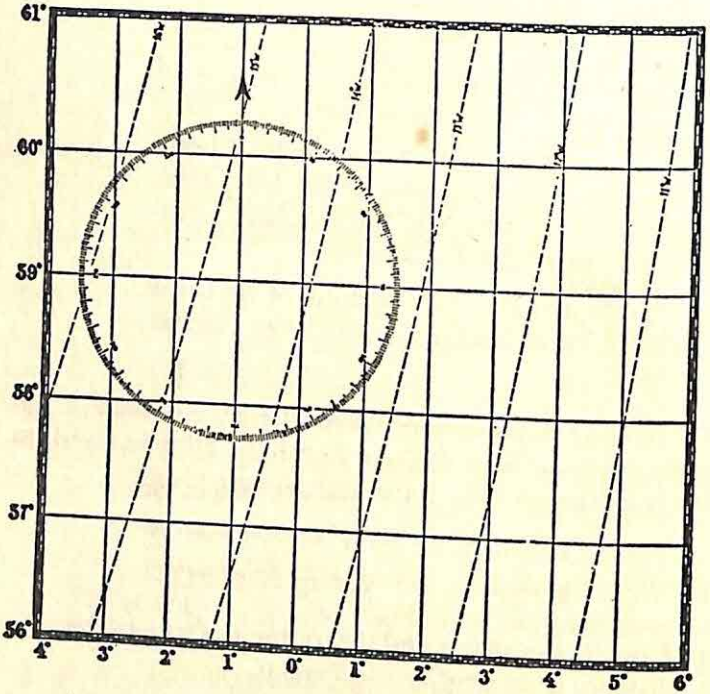
Leuchars	56° 17′ N 3° 00′ W
Bergen	60° 25′ N 5° 21′ E

1. Plot the two places and determine the true course and distance from Leuchars to Bergen.
2. If the compass heading and the deviation card are as given in the accompanying chart, what is the drift at the beginning of the flight?
3. How long would the flight take if a northeast wind were blowing?

[2] This drawing was made and the sample questions were written by F. J. Coyle of the Brooklyn Technical High School. In William A. Brownell *et al., The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, pp. 293–296). The University of Chicago Press, 1946.



FOR.	N	30	60	E	120	150
STEER	0	32	64	93	121	149
FOR.	S	210	240	W	300	330
STEER	180	209	238	267	295	329



Examiner's Scoring Check List for Navigation Problem

	Yes	*No*
1. Is W. longitude designated to the left of zero degrees and E. longitude to the right?		
2. Is the position of Leuchars plotted accurately within one minute?		
3. Is the position of Bergen plotted accurately within one minute?		

	Yes	No
4. What latitude scale did the student use to determine the distance from Leuchars to Bergen?		
a. 58° to 59°		
b. 56° to 61°		
c. 56° 17′ to 60° 25′		
5. Is the scaled distance from Leuchars to Bergen between 356 and 360 nautical miles? (Best answer, 358.7 nautical miles)		
6. Did the student specify nautical miles when stating distance?		
7. Is the true course from Leuchars to Bergen between 46° 00′ and 47° 00′? (Best answer, 46° 41′)		
8. Did the student read the compass heading as 68°?		
9. Did the student use the variation at the starting point, Leuchars?		
10. Did the student add the variation correction to true course to obtain a magnetic course?		
11. Did the student ascertain the deviation correction to be 4° West?		
12. Did the student add the deviation correction to the magnetic course to obtain compass course?		
13. Did the student subtract compass heading from compass course to determine the drift?		
14. Did the student ascertain the drift to be a West drift? (Best answer, 3° West drift)		
15. Did the student use ground speed in determining the time of flight?		

	Yes	*No*
16. Did the student correctly employ the wind triangle by		
a. laying off the drift angle to the right of true course?		
b. laying off the air speed along the true heading?		
c. assigning the wind vector a West, or left direction?		
d. using the same scale for measuring ground speed as was used in plotting air speed?		
e. obtaining a ground speed between 177 and 181 statute miles per hour? (Best answer, 179 statute miles per hour)		
17. Did the student use consistent units?		
18. Did the student divide the distance by the ground speed to obtain the time of flight?		
19. Was the time of flight between 2 hrs. 16 min. and 2 hrs. 20 min.? (Best answer, 2 hrs. 18 min.)		

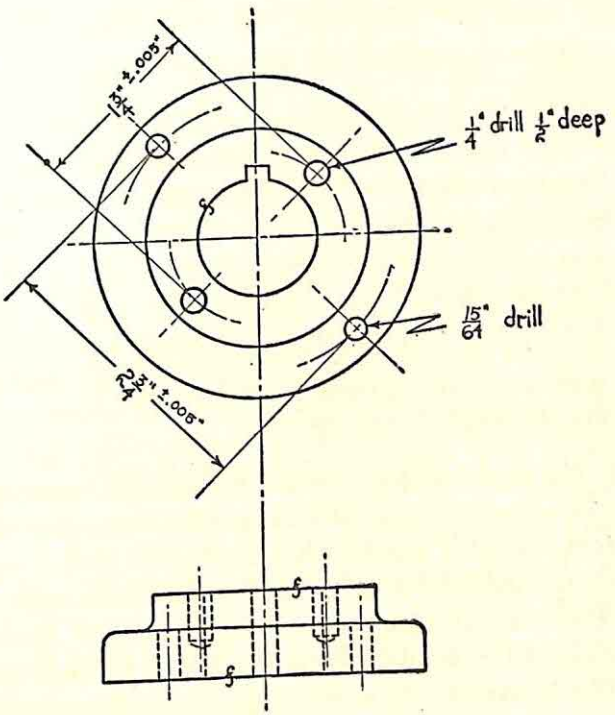
It should be noted that most of the items in the check lists above refer to numerical quantities and that when the student must base his answer on an estimate rather than on objectively established values, his answer must fall within a specified range in order to receive a score.

The check-list system of scoring is not limited to those problems in which the student calculates a numerical or algebraical answer. It can be used in all cases in which desirable behaviors can be specified in advance of administering the test. If such desirable behaviors cannot be specified, there is some question as to whether evaluations can be made at all.

The following example shows how a check list may be

used for evaluating solutions to a free-answer problem in which the student must perform some creative activity. It is given because of the common assumption that creative activity cannot be scored objectively.

Problem.[3] Design a simple drill jig for performing the last operation in the production of the part shown in the accompanying drawing.



The last operation is to "drill all holes." One thousand parts only are to be made and XLO bushings ½" O.D. and ½" long are to be used.

[3] This drawing was made and the sample questions were written by Emanuel Rosenthal of the Brooklyn Technical High School. In William A. Brownell *et al., The Measurement of Understanding* (45th Yearbook of the National Society for the Study of Education, Part I, pp. 299–300).

The following check list may be used for scoring the design which is evaluated in terms of the extent to which it will perform the required function. Such a check list may be employed diagnostically to indicate gaps in the student's understanding of the principles of design. It should be noted that the student is still given some freedom in designing the jig, provided that it will perform the required job satisfactorily.

Check List for Scoring the Design of the Jig	*Yes*	*No*
1. Did the student select the plug-type of jig?		
2. Did the student design the jig with a plug and key to fit the piece part?		
3. Does the jig fit against the finished surfaces of the piece part?		
4. Is it possible to reverse the jig and drill from both sides?		
5. Did the student design the correct bushing?		
6. Were the bushing holes located 90° apart with respect to the center line?		
7. Were the diameters dimensioned in decimals?		
8. Did the student insert a key in the plug to fit the piece part as a means of locking the work?		
9. Did the student draw both the top and the side views?		
10. Was a partial section drawn to show the bushing?		
11. Were the locations of the holes given in degrees?		
12. Did the student specify that the jig be made of C.R.S.?		

	Yes	No
13. Was the thickness of the jig plate equal to the height of the bushing?		
14. Was it specified that the jig be finished by grinding?		

DIFFICULTIES INHERENT IN THE PREPARATION AND SCORING OF ESSAY EXAMINATIONS

The general principles discussed in this appendix concerning the scoring of essay examinations provide only partial solutions to the many problems which arise in the preparation and use of essay-type examinations. Some of the problems and attempts to solve them are discussed below:

1. The Restriction of the Response. Many essay questions, because of their general vagueness, permit the student to expand his answer to any degree he sees fit. Under such conditions the essay may become a contest of who can write the most material in the given time, and two students may write different responses yet both may supply correct and appropriate information. Under such conditions scoring becomes an almost impossible task.

In order to overcome this difficulty it is necessary to inform the student about ways in which his response is to be restricted. It is usually desirable to limit his response to a given number of words, but more important still is it desirable to restrict his response in terms of the ideas he is to present. In other words, *essay questions should be specific and to the point, never* vague and general. *The more specific the question, the easier it is to score the response.*

The type of question which should be avoided and more appropriate types of questions on the same topics are given below:

Inappropriate Type of Question	More Appropriate Questions on the Same Topic
Discuss the causes of World War II.	List the main groups within Germany which supported Hitler in his rise to power.
Discuss the general situation in Europe in 1938.	Why did the allies fail to act when Hitler occupied the Rhineland?
	What excuse did Hitler give for invading Poland?
	How did the invasion of Ethiopia result in the crumbling of the system of collective security?
	What advantages did Russia gain by the 1939 pact with Germany?

The first group illustrates questions which are so general that the varied responses of the students cannot be scored on the same scale. The second group of questions restricts the response of students to limited domains but it would be wise also to restrict the response to a given length in each case.

If the response is adequately restricted by the nature of the question, the responses of students should differ quantitatively rather than qualitatively. This is easiest to explain by giving an example:

In response to an essay question, one student cites relevant facts A, B, C, D, E, F, while another student cites relevant facts A, B, C, D, E, F, G, H, I. The student who does best cites all the relevant facts, while the poorer student cites only some of these facts. There can be no doubt that

the one student does better than the other. In this case the responses differ quantitatively and the students may be compared in terms of their responses.

Suppose, on the other hand, that in response to a given question one student gives facts A, B, C, D, E, while another student gives facts T, U, V, W, X. In this case the two responses are so fundamentally different that there may be no valid basis for comparing them. The students are responding as if the stimulus situations were different in the two cases.

It is the latter situation that should be prevented in the construction of essay examinations and much can be done to avoid it by providing clear and unambiguous questions which restrict the response of the student to a considerable degree.

2. Limiting the Scoring System to Goals to be Measured. A major source of unreliability in the scoring of essay examinations is the common tendency to appraise factors other than those related to the goals of teaching. It has been clearly demonstrated that most teachers are influenced by such things as quality of handwriting, style, and organization even when these are not the outcomes which the examination is to appraise. The scorer may be largely unaware of the fact that he is influenced by these factors, particularly when he is basing his scores on a general impression rather than on a detailed analysis of the student's response. The conclusion to be drawn from this is that evaluation of essay tests should not be based on general impressions but should be made in terms of a scoring check list which can be used with some objectivity.

Sometimes English teachers insist that work in social studies, for example, should be scored for the quality of writing as well as for the objectives related to social studies. This is a good idea provided that separate scores are given

for the attainment of the English objectives and the social-studies objectives. Unless this is done the student is likely to be more confused than enlightened by his scores. He will not know whether a low score means inadequate achievement in English or in social studies or in both.

3. The Identity of the Responses and the Items in the Key. For scoring numerical problems it is not difficult to prepare a key which will correspond to the various parts of the response given by the students. However, when the response is non-numerical much greater difficulties are involved, for the restrictions placed on the response are usually considerably fewer. Consider, for example, the following essay question:

> Describe the main political events which occurred in Germany during the year in which Hitler became Chancellor.

If the check list were to list as one item the fact that President Hindenberg requested Hitler to form a cabinet, how is the student to be scored if he states that "the German President requested Hitler to form a cabinet," or "Hindenberg requested Hitler to form a cabinet"? The difficulty is to some extent but not entirely eliminated by breaking down the check list item into several components.

4. The Use of Essay Questions that Cannot be Scored Objectively. A considerable part of this discussion has been devoted to the use of essay examinations for measuring information. This is because they are mainly used for that purpose. The reader will undoubtedly have realized by this time that there are better ways of measuring information than are provided by the essay examination and that the main value of the essay as an evaluation device lies in other uses of it. There seems to be some merit in using the essay-type or free-response examination for measuring the ability

of the student to apply principles in solving problems, and the check-list system of scoring can be used in that connection. There are still other major uses of the essay, such as for measuring the ability of students to prepare compositions of literary worth,[4] the ability to prepare critical reviews, and the ability to prepare speeches which will serve a given purpose and have a particular appeal. These outcomes of education are considered important, but little is known as yet concerning the extent to which they can be measured by essay examinations of the traditional type. It is certainly not possible to score essays for these outcomes by using check lists of the kind which have so far been devised; therefore highly unreliable subjective scoring systems are the only ones available at the present time.

Attempts have been made to prepare English-composition rating scales in which scores are assigned separately for such things as punctuation, spelling, sentence structure, and organization, but such scales have been far from satisfactory.

5. Practicality of Scoring Essay Examinations with a Check List. The reader may rightly feel at this point that the tediousness of scoring essay examinations makes them impractical instruments where classes are large, the teaching loads heavy, and the teacher's time therefore limited. Under such circumstances the teacher would be wise to reserve the essay examination for measuring only those outcomes which cannot be measured by objective tests. Outcomes related to the acquisition of information and the acquisition of many skills can be measured much more effectively with objective-type tests than with free-answer tests.

[4] Since this book is limited to a discussion of tests that can be scored objectively, subjective systems of scoring cannot be discussed here. For a discussion of "The Measurement of Skill in Writing" the reader is referred to an article bearing that title by Paul B. Diederich and published in *The School Review,* December, 1946, pp. 584–592.

THE ESSAY AS A TEACHING DEVICE

This discussion of the essay has been concerned primarily with its uses as an evaluation instrument. This does not imply that it does not have other uses, and it must be recognized that it is a valuable teaching device. It can be used for giving the student opportunity for organizing his ideas, for expressing his thoughts, and for thinking out for himself the nature of important problems and their solutions. When the essay is used for the latter purpose many of the suggestions given here do not apply. For example, when the essay is used primarily as an exercise in thinking it may be undesirable to place too many restrictions on the response, but under such circumstances the responses of one student should not be compared with those of another.

A proper discussion of the use of the essay as a teaching device would occupy considerable space and is beyond the scope of this book. The paragraph above serves only to indicate that there are special problems involved in such use.

Index